РОССИЙСКАЯ АКАДЕМИЯ НАУК



ИНСТИТУТ ФИЛОСОФИИ

В. И. Шалак

ЛОГИЧЕСКИЙ АНАЛИЗ СЕТИ ИНТЕРНЕТ

Москва

Российская Академия Наук
Институт философии

В.И.Шалак

ЛОГИЧЕСКИЙ АНАЛИЗ СЕТИ ИНТЕРНЕТ

Москва
2005

УДК 681.142.37
ББК 32.817
Ш 18

В авторской редакции

Рецензенты:
доктор филос. наук *И.А.Герасимова*
кандидат филос. наук *В.О.Шангин*

Ш 18 **Шалак В.И.** Логический анализ сети Интернет. – М., 2005. – 96 с.

Монография посвящена вопросам анализа и построения логических моделей сети Интернет с целью более полного извлечения содержащейся в нем информации. Основной акцент делается на возможность извлечения не фактической информации, как это делается в настоящее время, а аналитической, явным образом не представленной в глобальной сети.

Книга предназначена для логиков, философов, специалистов по искусственному интеллекту и IT-технологиям, для исследователей, интересующихся возможностями применения точных методов в гуманитарных науках.

ISBN 5-9540-0047-6

СОДЕРЖАНИЕ

Предисловие 5
Аксиоматизация Интернет 8
Что мы будем понимать под сетью Интернет? 8
Что существенно для нашего анализа? 9
Логическая модель Интернет 10
Язык описания модели 16
Интерпретация 17
Аксиомы 18
Примеры использования языка 19
Анализ запросов поисковых систем 21
Алгебраическая модель 21
Об отношении логики и теории вероятностей 26
Вероятностная модель запросов 30
Подтверждение и принятие гипотез 33
Практический пример 1 38
Ряды событий 40
Практический пример 2 44
Практический пример 3 47
Практический пример 4 48
Математические методы контент-анализа 51
Что такое контент-анализ? 51
Оценки частот 53
Условные частоты 55
Нормы 56
Контекстный анализ 59
Связи категорий 60
Контент-мониторинг 61
Приложения 62
1. Комбинированная логика запросов 62
2. Алгоритм построения аналитических запросов . . 65
3. Технология прогноза 69
4. Летний банковский кризис 2004 года 82
Литература 94

ПРЕДИСЛОВИЕ

Мы привыкли к Интернету и обращение к нему для многих успело стать чем-то обыденным. Выйти в Интернет, просмотреть новостную ленту, получить и послать e-mail, заглянуть на форум, отыскать новую информацию по профессиональным интересам, разместить в сети что-то свое – для все большего числа людей эти действия превращаются в каждодневную рутину. Но обыденность Интернета обманчива. До сих пор нет единой точки зрения на то, что он есть такое?

Самая распространенная точка зрения заключается в том, что Интернет - это просто самая большая в мире электронная библиотека текстовой, графической, видео- и аудиоинформации практически по любым вопросам. Мы всегда можем подключиться к Интернету и посредством специальных поисковых систем извлечь из него необходимую нам информацию.

С другой точки зрения, Интернет – это некоторая новая реальность, которая предоставляет людям новые возможности по осуществлению политической, экономической, военной, культурной, научной и других видов деятельности. Президенты и правительства, промышленные и финансовые компании, военные и научные организации, учебные заведения, средства массовой информации и даже отдельные физические лица создают в сети Интернет свои представительства, вступают в определенные взаимоотношения друг с другом.

И уж совсем фантастическая точка зрения на Интернет как на материализовавшуюся ноосферу Вернадского, глобальную интеллектуальную систему, новую геологическую силу, которая в скором времени преобразит Землю до неузнаваемости. Не Интернет существует для людей, а мы в определенном смысле существуем для него и являемся всего лишь орудиями его познавательной деятельности. Эта точка зрения лишь кажется такой фантастической, но если присмотреться к сети Интернет повнимательней, то мы обнаружим, что он достаточно автономен, гибель любой его части не ведет к гибели всей системы, что в Интернете существуют активные центры, что в нем протекают процессы обмена информацией, одним из

следствий которых является усложнение и усовершенствование самой глобальной сети.

В настоящей работе мы будем рассматривать сеть Интернет как некоторое глобальное зеркало, которое распростерлось над реальным физическим миром и в котором тем или иным образом, с теми или иными искажениями отражаются события этого мира. Отдельные страницы всемирной сети – это всего лишь *пиксели* на *поверхности* зеркала, а сайты – небольшие группы пикселей. До сих пор, делая запросы к поисковым системам, мы интересовались содержанием отдельных пикселей, но не пытались получить глобальную картину того, что отражено в *зеркале*. В данном случае применимо выражение, что, взаимодействуя с сетью Интернет, мы *за деревьями не видели леса*. Интернет пока что является для нас источником фактов, а было бы хорошо, если бы он стал источником знаний.

Задача, которую мы перед собой ставим, может быть уточнена следующим образом. Пусть дана некоторая модель M_w, которая представляет реальный мир. Требуется построить модель M_i, представляющую Интернет, и определить, какие отношения между этими моделями имеют познавательную ценность, т.е. позволяют на основании свойств структуры M_i делать выводы о свойствах структуры M_w. Важность решения данной задачи состоит в том, что практически все содержание сети Интернет в полном объеме доступно каждому пользователю и требуется лишь научиться его анализировать. Если в физическом мире для уточнения параметров модели M_w нам зачастую приходится проводить ресурсоемкие исследования, то, изучая модель M_w посредством анализа модели M_i, мы практически не расходуем никаких ресурсов. Понятно, что не всякий элемент структуры M_w дублирован в M_i и доступен такого рода анализу, но даже то, что находит отражение в Интернет, все равно поражает своим объемом.

В настоящее время существует направление исследования Интернет, получившее название *web-mining*. Однако круг задач, которые решают в его рамках, в основном ограничен вопросами эффективного поиска, категоризацией текстов, изучением траекторий, по которым перемещаются пользователи глобальной сети, кликая мышкой по гипертекстовым ссылкам. Задачи

интересные, но чисто утилитарные, так как преследуют цель улучшения существующих подходов, а не выход за их рамки.

В числе вопросов, на которые может дать ответ логический анализ, следующие:

1. Какие *типы данных* используются в модели M_i для представления информации о модели M_w?
2. Как представлено *время* в M_i и как оно соотносится с временем M_w?
3. Что есть *событие* в модели M_i?
4. Что значит *существовать* в M_i?
5. Проблема *истинности* в M_i, и ее отношение к истинности в M_w?
6. Каковы *методы рассуждений* над M_i?
7. Каковы *методы поиска закономерностей* в M_i?
8. Возможно ли построение *баз знаний* над M_i?
9. Как *распространяется информация* в M_i?

Полагаем, что приведенный перечень вопросов не является исчерпывающим. Для ответа на них потребуются усилия многих исследователей, но и результат будет стоить того. В настоящей книге мы коснемся лишь части из них, оставив другие для будущих более детальных и глубоких исследований.

АКСИОМАТИЗАЦИЯ ИНТЕРНЕТ

Что мы будем понимать под сетью Интернет?

На самом низком физическом уровне Интернет представляет из себя просто большое число компьютеров, соединенных между собой посредством электрических проводов, оптоволоконных кабелей, каналов радиосвязи и пр. Особого интереса данная структура для логиков не представляет, так как речь идет всего лишь о способе ее технической реализации *в железе*.

На более высоком уровне Интернет состоит не из компьютеров, а из серверов, основная функция которых заключается в хранении информации и ее передаче по определенным правилам (протоколам) другим серверам. Для логиков определенный интерес может представлять анализ протоколов обмена информацией. Здесь находит применение аппарат многосубъектных эпистемических логик. Могут решаться задачи определения логической корректности протокола. Известно, что многие протоколы (наборы правил) обмена информацией между серверами содержат ошибки, которые при определенных условиях могут приводить к некорректной работе. Знание этих недостатков позволяет злоумышленниками получать несанкционированный доступ к различным информационным системам, имеющим связь с Интернет. Логический анализ и устранение таких недостатков является интересной, но все-таки частной задачей.

На еще более высоком уровне, к которому мы собственно и привыкли, Интернет представляет из себя множество сайтов, состоящих в свою очередь из страниц, на которых может быть размещена текстовая, графическая, видео и аудиоинформация. На страницах имеются ссылки, связывающие их с другими страницами и сайтами, что в конечном счете образует гипертекстовую структуру, получившую официальное название World Wide Web – Всемирная Паутина.

Именно последний уровень представления Интернета и будет нас интересовать.

Что существенно для нашего анализа?

Интернет развивается очень бурно. Постоянно совершенствуются способы представления информации на Интернет-страницах, расширяются старые и возникают новые языки для их кодирования. Проблема представления информации также имеет прямое отношение к логике, но в данной работе нас будет интересовать не она. Мы предполагаем, что информация уже тем или иным образом представлена, и задача, которая стоит перед нами, - научиться эффективно пользоваться этой информацией. Поэтому мы отвлечемся от конкретных решений и их реализаций и постараемся принять более общую точку зрения, которая менее подвержена изменениям, связанным с эволюцией Интернет. Нам важно не увязнуть в сиюминутных деталях, а получить результаты, которые останутся значимы еще долгое время.

Более общая точка зрения заключается в том, что Интернет – это реляционная структура, элементарным типом которой являются цепочки символов. Всякая страница сети Интернет – это просто цепочка символов, подчиняющаяся определенному синтаксису. Если мы хотим создать Интернет-страницу, мы должны всего лишь составить некоторый текст и сохранить его на специальном компьютере, подсоединенном к глобальной сети Непосредственно на странице хранится лишь текстовая информация, а графическая, видео и аудиоинформация представлены специальными ссылками на файлы соответствующего формата. Ссылки – это тоже цепочки символов. Специальные программы – интерпретаторы языков, с помощью которых закодированы Интернет-страницы, знают, как найти по ссылкам нужные файлы и представить пользователю в удобном виде содержащуюся в них информацию.. Как это конкретно делается в каждом отдельном случае, для нас совершенно неважно. Важно лишь, что это делается и всегда будет делаться.

Кроме четырех упомянутых выше видов информации в Интернете широко представлена также алгоритмическая информация. Когда мы набираем текст запроса в поисковой системе и нажимаем на кнопку «*Поиск*», мы задействуем алгоритмическую информацию. Некоторые сайты специализируются именно на ней. Описания алгоритмов, которые

при этом используются, также либо закодированы в самой странице, либо представлены ссылками на соответствующие файлы.

Мы принимаем в качестве базового типа данных сети Интернет цепочки символов - слова в определенном алфавите. Базовые операции с ними нам хорошо знакомы. Все остальные, более сложные, типы данных мы должны будем определить с их помощью.

Логическая модель Интернет

Для того чтобы появились цепочки символов, мы должны зафиксировать начальный алфавит букв Alpha, из которых эти цепочки будут строиться. Чтобы не слишком отрываться от действительности, будем считать, что множество букв Alpha конечно. Одним из примеров такого алфавита является хорошо знакомый набор из 256 ASCII-символов. Над этим алфавитом определим множество слов Word:

Def.1
1. Если $a \in Alpha$, то $a \in Word$;
2. Если $v \in Word$ и $w \in Word$, то $vw \in Word$;
3. Ничто другое словом не является.

Базовым отношением на множестве $Word \times Word$ является отношение вхождения Include слова v в слово w, которое определяется очевидным образом:

Def.2 $Include \subset Word \times Word$, удовлетворяющее условию
- $\langle v,w \rangle \in Include \Leftrightarrow \exists x,y \in Word$(w=v или w=xv или w=vy или w=xvy)

Мы могли бы определить и другие известные типы отношений и операций над словами, но не станем этого делать, так как их добавление ничего принципиально нового не дает. Важно лишь иметь ввиду, что любые наши действия в конечном счете всегда сводимы к базовым операциям со словами в некотором фиксированном алфавите Alpha.

Мы знаем, что всякое физическое тело имеет пространственно-временные координаты. Нечто подобное

свойственно и Интернет. В нем также имеются свои *тела* - Интернет-страницы как слова в алфавите Alpha, построенные в соответствии с синтаксисом языка HTML или его модификаций.

Def.3 Body $\subset$ Word

Никаких ограничений на размер данного множества мы не налагаем. Важно лишь то, что мы всегда можем эффективно определить, принадлежит некоторое слово b множеству Body или не принадлежит. Это означает, что множество Body рекурсивно.

Как и у физических тел, у Интернет-страниц есть свои *координаты* в пространстве глобальной сети. В качестве координат для пользователей Интернет выступают построенные по определенным правилам URL-адреса страниц, также являющиеся словами в нашем алфавите.

Def.4 Address $\subset$ Word

На размер этого множества мы также не налагаем никаких ограничений и предполагаем лишь рекурсивность.

Заметим, что далеко не каждому элементу множества Address, соответствует реально существующая страница. Пользователям Интернет знакома «*Ошибка 404. Файл не найден*». Это сообщение как раз и говорит о том, что была совершена неудавшаяся попытка перейти по адресу, которому не соответствует ни одна реально существующая страница. В физическом мире тоже не все места в пространстве заняты телами, встречается и пустота.

Помимо этого каждой странице сопоставлено *время* ее создания или последней модификации. Реализуется оно через систему временных меток, которые также являются словами в алфавите Alpha.

Def.5 Time $\subset$ Word.

Множество Time рекурсивно и на нем задано рекурсивное отношение линейного порядка, которое будем обозначать посредством $<$.

Аналогия между физическими телами и Интернет-страницами может быть продолжена. Как и физические тела, страницы глобальной сети *взаимодействуют* друг с другом. Воздействие происходит через посредство ссылок (адресов), и благодаря этому World Wide Web приобретает гипертекстовую структуру и связность.

Интернет-страница появляется тогда, когда некоторый код страницы (тело) размещается по определенному адресу. Это позволяет дать следующее определение:

Def.6 $Page \subset Address \times Body \times Time \times 2^{Address}$, удовлетворяющее условиям:

- $\langle a,b1,t1,R1\rangle \in Page$ и $\langle a,b2,t2,R2\rangle \in Page \Rightarrow b1=b2,\ t1=t2,\ R1=R2$ – функциональность отношения, т.е. страница однозначно задается ее адресом;
- $\langle a,b,t,R\rangle \in Page\ \&\ r \in R \Rightarrow \langle r,b\rangle \in Include$;
- Page конечно.

Следующая интересующая нас структура – это *сайт*, некоторое конечное множество страниц. Сайты характеризуются тем, что у них есть одна так называемая главная страница, адрес которой считается адресом самого сайта.

Def.7 $Site \subset Page \times 2^{Page}$, удовлетворяющее условиям:

- $\langle p,P1\rangle \in Site$ и $\langle p,P2\rangle \in Site \Rightarrow P1=P2$ - функциональность;
- $\langle p,P\rangle \in Site \Rightarrow p \in P$ – главная страница сайта принадлежит самому сайту;
- $\langle p1,P1\rangle \in Site$ и $\langle p2,P2\rangle \in Site\ \&\ p1 \neq p2 \Rightarrow P1 \cap P2 = \varnothing$ - одна и та же страница не может принадлежать одновременно двум сайтам;
- $\{p \mid \exists x \exists P(\langle x,P\rangle \in Site \& p \in P)\} = Page$ – каждая реально существующая страница принадлежит хотя бы одному из сайтов.

Так как каждая страница идентифицируется по адресу в Интернет, возможно альтернативное определение сайтов:

Def.7' $Site \subset Address \times 2^{Address}$, удовлетворяющее условиям:

- $\langle a,A1\rangle \in Site$ и $\langle a,A2\rangle \in Site \Rightarrow A1=A2$ - функциональность;
- $\langle a,A\rangle \in Site \Rightarrow a \in A$ – главная страница сайта принадлежит самому сайту;

- $\langle a1,A1\rangle \in Site$ и $\langle a2,A2\rangle \in Site$ & $a1 \neq a2 \Rightarrow A1 \cap A2 = \varnothing$ - одна и та же страница не может принадлежать одновременно двум сайтам;
- $\{a \mid \exists x \exists A(\langle x,A\rangle \in Site \& a \in A)\} = \{a \mid \exists b \exists t \exists R Page\langle a,b,t,R\rangle \in Page\}$ – каждая реально существующая страница принадлежит хотя бы одному из сайтов.

Именно это определение мы и будем использовать в дальнейшем.

Еще одной структурой Интернет, на которую мы хотим обратить внимание, являются домены. Они позволяют объединять различные сайты в тематические группы, задавая на них древовидный порядок. О принадлежности сайта к тому или иному домену можно судить по его адресу, так как составными частями адреса являются имена доменов. В нашем представлении домен – это пара, состоящая из имени домена и множества сайтов, которые ему принадлежат. Внутреннюю структуру имен доменов мы анализировать не будем.

Def. 8 $Domain \subset Word \times 2^{Address}$, удовлетворяющее условиям

- $\langle n,A1\rangle \in Domain$ & $\langle n,A2\rangle \in Domain \Rightarrow A1 = A2$ – функциональность;
- $\langle n,A1\rangle \in Domain$ & $\langle m,A2\rangle \in Domain \Rightarrow A1 \cap A2 = \varnothing$ или $A1 \cap A2 = A1$ – множества сайтов, принадлежащие любым двум доменам либо дизъюнктны, либо одно из них является подмножеством другого;
- $\cup\{a \mid \exists w \exists d(\langle w,d\rangle \in Domain \& a \in d)\} = \{a \mid \exists x(\langle a,x\rangle \in Site)\}$ – любой сайт принадлежит хотя бы одному домену.

И наконец последним элементом нашей модели Интернет являются поисковые системы. Не будь их, каждому пользователю был бы доступен лишь ограниченный крохотный фрагмент глобальной сети. Именно создателям поисковых систем мы должны быть благодарны за то, какую роль стала играть сеть Интернет в нашей жизни. В ответ на запрос, сформулированный в специальном языке, поисковая система возвращает некоторое конечное множество адресов Интернет-страниц, удовлетворяющих условиям запроса, с указанием на время, когда они были проиндексированы, т.е. занесены в базу данных поисковой системы. База данных поисковой системы является как бы ее внутренним представлением Интернет. Важной

особенностью это базы данных является то, что она принципиально неполна, так как в глобальной сети постоянно появляются новые страницы, но не все они и не сразу заносятся в базу. Одновременно идет и противоположный процесс – страницы исчезают из всемирной паутины, но упоминание о них все еще хранится в базе.

Для начала нам необходимо определить множество слов-запросов Request, посредством которых пользователь дает поисковой системе задание найти те или иные страницы.

Def.9

1. Если w –слово в алфавите Alpha-{ ∧,#,-,),(}, то $w \in$ Request;
2. $w \in$ Request и $v \in$ Request $\Rightarrow$ $(w \wedge v) \in$ Request, $(w\#v) \in$ Request, $(w \wedge -v) \in$ Request;
3. Ничто иное словом-запросом не является.

В качестве образца мы взяли языки запросов таких поисковых систем Интернет как AltaVista, Rambler и Яndex. Интересно обратить внимание на используемый в них язык. В нем присутствуют связки конъюнкции ∧, дизъюнкции # и отрицания -. При этом на использование отрицания налагается ограничение. Его можно использовать лишь вместе с конъюнкцией. Т.е. фактически используется не само отрицание, а в язык вводится третья бинарная связка ∧- со смысловой интерпретацией '... и не ...', которая в классической логике выражает то же самое, что и отрицание импликации. Интересной особенностью данного языка является то, что в нем невозможно выразить универсально значимое высказывание, т.е. невозможно сформулировать такой запрос, ответом на который было бы множество ссылок на все проиндексированные в поисковой системе страницы.

Определение поисковой системы будет выглядеть следующим образом:

Def.10 SE $\subset$ Request×Address×Time, для которого дополнительно выполняются условия:

- $\langle q,a,t1\rangle \in SE \,\&\, \langle q,a,t2\rangle \in SE \Rightarrow t1=t2$ – в базе данных поисковой системы хранится лишь время последней модификации страницы;

- $\langle q,a,t\rangle \in SE \;\&\; \neg\exists g \exists h(q=(g \wedge h) \text{ или } q=(g\#h) \text{ или } q=(g \wedge -h)) \Rightarrow \exists b \exists R(\langle a,b,t,R\rangle \in Page \text{ и } \langle q,b\rangle \in Include)$;
- $\langle (q \wedge g),a,t\rangle \in SE \Leftrightarrow \langle q,a,t\rangle \in SE \text{ и } \langle g,a,t\rangle \in SE$;
- $\langle (q\#g),a,t\rangle \in SE \Leftrightarrow \langle q,a,t\rangle \in SE \text{ или } \langle g,a,t\rangle \in SE$;
- $\langle (q \wedge -g),a,t\rangle \in SE \Leftrightarrow \langle q,a,t\rangle \in SE \text{ и } \langle g,a,t\rangle \notin SE$.

Иногда вместо $\langle q,a,t\rangle \in SE$ мы будем писать $\langle a,t\rangle \in SE(q)$.

Технология поиска информации в сети Интернет постоянно развивается. Создаются новые более совершенные языки запросов, помогающие точно отыскивать ту информацию, которая действительно интересует пользователя. Анализ языков запросов и его совершенствование – интересная задача, решению которой могли бы помочь логики-философы. В настоящее время результатом запроса к поисковой системе является набор фактов, удовлетворяющих заданным условиям. Это происходит потому, что для непосредственных разработчиков поисковых систем Интернет – это просто некоторая структура хранения и обмена информации. Для логика-философа Интернет – это некоторое отражение явлений и процессов реального физического мира. С определенными модификациями все философские категории, которые создавались для упорядочивания окружающего нас мира, могут быть перенесены и замечены в структуре глобальной сети. Ориентироваться во внешнем мире нам помогает знание закономерностей, а не фактов. Будущее не за поисковыми, а за поисково-аналитическими системами, результатом взаимодействия с которыми будут не только факты, но и новые знания.

Пока же мы ограничимся лишь тем, что уже есть. Для простоты будем считать, что в Интернете существует всего одна поисковая система, которой все и пользуются.

Итак, в каждый момент времени модель Интернет можно представить в виде

$$M_i = \langle Alpha, Word, Include, Address, Body, Time, Page, Site, Domain, Request, SE\rangle$$

Язык описания модели

Определим язык, с помощью которого мы сможем формулировать утверждения о свойствах Интернет. Мы хотим ограничиться первопорядковым языком всего лишь с одним сортом переменных.

Def.11 Исходные символы языка

1. Множество констант Letter;
2. Множество индивидных переменных Var ;
3. Двухместные функциональные символы *, ∧, #, ∧-;
4. Одноместные предикатные символы Address, Body, Time, Request;
5. Двухместные предикатные символы =, In, Site,Domain;
6. Трехместный предикатный символ SE;
7. Четырехместный предикатный символ Page;
8. Логические связки &, ¬;
9. Кванторы ∃;
10. Скобки (,).

Def.12 Термы

1. Всякая константа $c \in$ Letter есть терм;
2. Всякая переменная $v \in$ Var есть терм;
3. Если t1 и t2 – термы, то (t1*t2), (t1∧t2), (t1#t2), (t1∧-t2); – термы;
4. Ничто другое термом не является.

Def.13 Формулы

1. Если t1, t2, t3, t4 – термы, то Address(t1), Body(t1), Time(t1), Request(t1), t1=t2, In(t1,t2), Site(t1,t2), Domain(t1,t2), SE(t1,t2,t3) , Page(t1,t2,t3,t4) – формулы;
2. Если $x \in$ Var, а A и B – формулы, то (A&B), ¬A , ∃xA – формулы;
3. Ничто другое формулой не является.

Связки ⊃, ≡, ∨ и квантор ∀ вводим обычным образом с помощью определений.

Интерпретация

Определим функцию интерпретации F, которая будет сопоставлять исходным нелогическим символам нашего языка различные объекты модели Mi.

Def.14

1. $F(c)\in Alpha$, $c\in Letter$
2. $F(*):Word\times Word\to Word = \{\langle w,v,wv\rangle \mid w\in Word, v\in Word\}$
3. $F(Address)=Address$
4. $F(Body)=Body$
5. $F(Time)=Time$
6. $F(Request)= Request$
7. $F(In)=Include$
8. $F(Site)=\{\langle a,b\rangle \mid \exists A(\langle a,A\rangle\in Site \;\&\; b\in A)\}$
9. $F(SE)=SE$
10. $F(Domain)=\{\langle n,a\rangle \mid \exists A(\langle n,A\rangle\in Domain \;\&\; a\in A)\}$
11. $F(Page)=\{\langle a,b,t,r\rangle \mid \exists R(\langle a,b,t,R\rangle\in Page \;\&\; r\in R)\}$

Для фиксированной модели Mi, функции интерпретации F и приписывания значений индивидным переменным $v\in Val = Word^{Var}$ определим значение терма t следующим образом:

Def.15

1. Если $c\in Letter$, то $F^v(c)=F(c)$;
2. Если $x\in Var$, то $F^v(x)=v(x)$;
3. Если t1, t2 – термы, то $F^v(t1*t2)=F^v(t1)F^v(t2)$, $F^v(t1\wedge t2)=(F^v(t1)\wedge F^v(t2))$, $F^v(t1\#t2)=(F^v(t1)\#F^v(t2))$, $F^v(t1\wedge\text{-}t2) = (F^v(t1)\wedge\text{-}F^v(t2))$.

Отношение $\langle Mi,F,v\rangle\models A$ – «формула A истинна в модели Mi при интерпретации F и приписывании v» определяется обычным образом.

Def.16

1. $\langle Mi,F,v\rangle\models t1=t2 \Leftrightarrow F^v(t1)= F^v(t2)$
2. $\langle Mi,F,v\rangle\models Address(t) \Leftrightarrow F^v(t)\in F(Address)$;
3. $\langle Mi,F,v\rangle\models Body(t) \Leftrightarrow F^v(t)\in F(Body)$;
4. $\langle Mi,F,v\rangle\models Time(t) \Leftrightarrow F^v(t)\in F(Time)$;
5. $\langle Mi,F,v\rangle\models Reques(t) \Leftrightarrow F^v(t)\in F(Request)$;
6. $\langle Mi,F,v\rangle\models In(t1,t2) \Leftrightarrow \langle F^v(t1),F^v(t2)\rangle\in F(In)$;
7. $\langle Mi,F,v\rangle\models Site(t1,t2) \Leftrightarrow \langle F^v(t1),F^v(t2)\rangle\in F(Site)$;

8. <Mi,F,v>|=SE(t1,t2,t3) ⇔ < F^v(t1),F^v(t2),F^v(t3)>∈F(SE);
9. <Mi,F,v>|=Domain(t1,t2) ⇔ <F^v(t1),F^v(t2)>∈F(Domain);
10. <Mi,F,v>|=Page(t1,t2,t3,t4) ⇔ <F^v(t1),F^v(t2),F^v(t3),F^v(t4)>∈F(Page);
11. <Mi,F,v>|=(A&B) ⇔ <Mi,F,v>|=A и <Mi,F,v>|=B ;
12. <Mi,F,v>|=¬A ⇔ неверно, что <Mi,F,v>|=A ;
13. <Mi,F,v>|=∃xA ⇔ для некоторого v', отличного от v возможно лишь значением, приписываемым переменной x, имеет место <Mi,F,v'>|=A.

Отношения «<Mi,F>|=A - формула A истинна в модели Mi при интерпретации F» и «формула |=A общезначима» также определяются обычным образом:

Def.17 <Mi,F>|=A ⇔ для всякого приписывания v∈Val имеет место <Mi,F,v>|=A

Def.18 |=A ⇔ для всякой Интернет-модели Mi и всякой интерпретации F имеет место <Mi,F>|=A

Привычным образом расширим язык логическими связками ∨, ⊃,≡ и квантором ∀.

Ниже приведен набор аксиом, которые общезначимы в наших моделях. Этот набор не является семантически полной системой аксиом хотя бы потому, что в модели часть предикатов конечны, а свойство конечности, как известно, невыразимо в первопорядковой логике.

Аксиомы

1. t1*(t2*t3)=(t1*t2)*t3
2. In(x,x)
3. In(v,w) ⊃ ∃x∃y(w=v ∨ w=x*v ∨ w=v*y ∨ w=x*v*y)
4. Page(a,b,t,r) ⊃ Address(a)&Body(b) &Time(t)&Address(r)
5. Page(a,b1,t1,r1)&Page(a,b2,t2,r2) ⊃ (b1=b2&t1=t2)
6. Page(a,b,t,r) ⊃ In(r,b)
7. Site(m,a) ⊃ Address(m)&Address(a)
8. Site(m,a) ⊃ Site(m,m)
9. Site(m,a)&Site(n,a) ⊃ m=n
10. Site(m,a) ⊃ ∃b∃t∃rPage(a,b,t,r)
11. Page(a,b,t,r) ⊃ ∃xSite(x,a)

12. $Domain(n,a) \supset Address(n) \& Address(a)$

13. $Domain(m,a1) \& Domain(n,a2) \supset \neg\exists y(Domain(m,y) \& Domain(n,y)) \vee \forall y(Domain(m,y) \supset Domain(n,y))$

14. $\exists x Domain(x,y) \equiv \exists z Site(y,z)$

15. $SE(q,a,t) \supset Request(q) \& Address(a) \& Time(t)$

16. $SE(q,a,t1) \& SE(q,a,t2) \supset t1=t2$

17. $SE(q,a,t) \& \neg\exists g \exists h(q=(g \wedge h) \vee q=(g \# h) \vee q=(g \wedge -h)) \supset \exists b \exists r(Page(a,b,t,r) \& In(q,b))$

18. $SE((q \wedge g),a,t) \equiv SE(q,a,t) \& SE(g,a,t)$

19. $SE((q \# g),a,t) \equiv SE(q,a,t) \vee SE(g,a,t)$

20. $SE((q \wedge -g),a,t) \equiv SE(q,a,t) \& \neg SE(g,a,t)$

Примеры использования языка

Утверждение «на странице по адресу http://www.iph.ras.ru/~logic/index.html имеется фраза *Сектор логики ИФРАН*» в нашем языке может быть записано следующим образом:

$$\exists b \exists t \exists r(Page(\text{‘http://www.iph.ras.ru/\textasciitilde logic/index.html’},b,t,r) \& In(\text{‘Сектор логики ИФРАН’},b))$$

Легко проверить, что данное утверждение истинно (было истинно в момент написания работы). Если данная страница проиндексирована в поисковой системе, то имеет место <‘http://www.iph.ras.ru/~logic/index.html’,t>∈SE(‘Сектор логики ИФРАН’) для некоторого момента времени t, когда эта страница была занесена в базу данных поисковой системы.

Также легко проверить истинность более сложного утверждения

$$\exists b \exists t \exists r(Page(\text{‘http://www.iph.ras.ru/\textasciitilde logic/staff.html’},b,t,r) \& In(\text{‘Карпенко’},b) \& In(\text{‘Анисов’},b) \& In(\text{‘Васюков’},b) \& (In(\text{‘Шкатов’},b) \vee In(\text{‘Зиновьев’},b)))$$

Поэтому, если задать запрос в форме ‘Карпенко∧Анисов∧Васюков∧(Шкатов#Зиновьев)’, то есть шанс среди ответов получить адрес ‘http://www.iph.ras.ru/~logic/staff.html’.

Когда пользователи Интернет ищут в нем какую-либо информацию, они интересуются областями истинности тех или иных формул. Результатом такого поиска, в случае его успешности, является информация фактического характера. Для ее представления, с точки зрения логики, достаточно бескванторной логики предикатов, не содержащей переменных, а лишь одни константы. Это слишком бедный язык, чтобы на нем и остановиться.

Предложенная модель и система аксиом не являются истинами в последней инстанции и допускают многие уточнения и улучшения. Структура страниц далеко не так проста, как показано выше. Для решения некоторых задач необходимо более детальное ее представление. Не так проста структура сайтов, структура адресов и доменов. Поисковые системы вообще представляют богатейшее поле для практических приложений логики.

АНАЛИЗ ЗАПРОСОВ ПОИСКОВЫХ СИСТЕМ

Если принять точку зрения, что Интернет – это некоторое зеркало, в котором находят отражение события реальной земной жизни, которые вовсе не случайны, а связаны между собой различными закономерностями, то следы этих закономерностей должны присутствовать и в глобальной сети. Нужно просто их обнаружить и извлечь. Если это окажется возможным, то мы получим богатейший ресурс – сможем изучать окружающий мир, анализируя его отражение в зеркале Интернета, который постоянно наполняется новым содержанием и доступен каждому.

Всякая закономерность – это запрет на осуществление определенных состояний. Чем сильнее запрет, тем сильнее закономерность. Если запреты отсутствуют, то могут реализоваться любые возможные состояния, и энтропия такой среды максимальна. Интернет является информационной средой. Для поиска закономерностей, мы должны искать в нем неравномерность распределения информации. До сих пор поиск ограничивался анализом содержания отдельных страниц. Если использовать терминологию многомерных пространств, это лишь одно из измерений Интернета. В предлагаемой модели присутствуют и другие измерения – *адреса, временной порядок, взаимные ссылки, распределение информации по сайтам и доменам и пр.*

Алгебраическая модель

Нам потребуется ввести некоторые дополнительные обозначения, так как в дальнейшем изложении мы будем различать слово-запрос, сам акт запроса и ответ на него. Во-первых, что такое запрос к поисковой системе? Имеется слово w, которое мы передаем поисковой системе SE для выполнения определенных действий. Под запросом к поисковой системе будем понимать сам акт взаимодействия с нею по определенным правилам. Для его обозначения мы будем использовать выражения вида R(w). В качестве реакции на запрос поисковая система SE приступает к поиску в своей базе данных адресов Интернет-страниц, которые *удовлетворяют* слову-запросу w. Условия удовлетворения слову-запросу w представлены в определении Def.10 и аксиомами 17-20.

По окончании поиска в ответ на запрос R(w) формируется множество адресов страниц, удовлетворяющих слову-запросу w. Его мы запишем в виде $R[w] = \{\langle a,t\rangle | \langle w,a,t\rangle \in SE\}$ и будем называть *областью истинности запроса*. Очевидно, что R[w] имеет объемную характеристику и количественную, так как Интернет является конечной структурой. Для обозначения количественной характеристики ответа на запрос R(w) мы будем использовать запись |R[w]| - мощность множества $R[w] = \{\langle a,t\rangle | \langle w,a,t\rangle \in SE\}$.

Запишем в терминах теоретико-множественных операций над областями истинности, как выглядят ответы на различные типы запросов.

Если w и v - два слова-запроса, то ответами на запросы $R(w \wedge v)$, $R(w \vee v)$ и $R(w \wedge \neg v)$ будут следующие множества:

- $R[w \wedge v] = \{\langle a,t\rangle | \langle w,a,t\rangle \in SE$ и $\langle v,a,t\rangle \in SE\} = \{\langle a,t\rangle | \langle w,a,t\rangle \in SE\} \cap \{\langle a,t\rangle | \langle v,a,t\rangle \in SE\} = R[w] \cap R[v]$

- $R[w \vee v] = \{\langle a,t\rangle | \langle w,a,t\rangle \in SE$ или $\langle v,a,t\rangle \in SE\} = \{\langle a,t\rangle | \langle w,a,t\rangle \in SE\} \cup \{\langle a,t\rangle | \langle v,a,t\rangle \in SE\} = R[w] \cup R[v]$

- $R[w \wedge \neg v] = \{\langle a,t\rangle | \langle w,a,t\rangle \in SE$ и $\langle v,a,t\rangle \notin SE\} = \{\langle a,t\rangle | \langle w,a,t\rangle \in SE\} \cap \{\langle a,t\rangle | \langle v,a,t\rangle \notin SE\} = R[w] - R[v]$

С последним пунктом не все так просто, как может показаться на первый взгляд. Дело в том, что невозможно сформировать запрос, областью истинности которого было бы множество $\{\langle a,t\rangle | \langle v,a,t\rangle \notin SE\}$. Т.е. пользователю недоступно обращение к универсальному множеству $\{\langle a,t\rangle | \exists w(\langle w,a,t\rangle \in SE\}$ для взятия дополнения относительно него. Дополнение всегда берется относительно другого уже сформированного множества. В нашем конкретном случае это было множество $\{\langle a,t\rangle | \langle w,a,t\rangle \in SE\}$. В качестве аналогии можно привести аксиому выделения теории множеств Цермело-Френкеля. В ней также не существует универсальных множеств, и все операции над ними ограничены ранее построенными множествами.

Сказанное выше означает, что множество ответов на запросы представляет хороший пример ультраинтуиционистской структуры [2], с которой мы, оказывается, сталкиваемся

буквально каждый день. Семейство множеств 2^{SE} конечно и образует решеточную структуру, так как любые два его элемента имеют пересечение и объединение. Эта решетка имеет наименьший элемент, но не имеет наибольшего (которым должно быть и на самом деле является множество SE), а потому не образует булевой алгебры. Мы знаем, что конечная решетка всегда имеет наибольший элемент, но в том-то и сложность ситуации, что мы имеем дело с решеткой, которая существует не сама по себе, а для некоторого внешнего наблюдателя, который не способен идентифицировать наибольший элемент. В этом и только этом смысле наибольший элемент не существует. Субъект *воспринимает* 2^{SE} *как бесконечную решетку с конечными объединениями и пересечениями*.

Простейший запрос с использованием отрицания имеет вид R(w∧-v), для которого R[w∧-v] = {<a,t>| <w,a,t>∈SE и <v,a,t>∉SE). Т.е. для правильного использования отрицания мы должны указать область соотнесения, относительно которой берется дополнение. В терминах нормальных форм классической логики высказываний правильно построенными запросами являются те, которые представимы в виде дизъюнктивной нормальной формы (дизъюнкции элементарных конъюнкций), где каждая элементарная конъюнкция содержит хотя бы один конъюнкт без отрицания. Легко показать, что для всякого правильного построенного запроса R(w) семейство множеств $2^{R[w]}$ образует булеву алгебру, наибольшим элементом которой является R[w].

Ограничение на использование отрицания ни в коем случае не является существенным, так как фактически требует от пользователя поисковых систем всего лишь определить универсум, относительно которого будет браться дополнение. Своеобразная дисциплина мышления.

Посмотрим теперь, можно ли чисто аналитически вычислить количественные оценки |R[w∧v]|, |R[w∨v]| и |R[w∧-v]| по оценкам |R[w]| и |R[v]|?

Так как R[w∧v] = R[w]∩R[v], то для произвольных w и v оценка |R[w∧v]| может принимать любые из значений в интервале от 0 до min(|R[w]|, |R[v]|). Нулевую оценку мы получим, когда два данных множества имеют пустое пересечение, а оценку |R[w∧v]|

$= \min(|R[w]|,|R[v]|)$ получим, когда одно из множеств включено в другое. Двойственным образом оценивается ответ на запрос вида $R(w \vee v)$, ответом на который будет $R[w \vee v] = R[w] \cup R[v]$. Минимальным значением, которое может принять $|R[w \vee v]|$, является $\max(|R[w]|, |R[v]|)$, когда одно из множеств включено в другое, а максимальным является $|R[w]|+|R[v]|$, когда два множества дизъюнктны. Оценки $|R[w \wedge -v]|$ заключены в интервале от 0 до $|R[w]|$ и могут быть вычислены по формуле что $|R[w \wedge -v]| = |R[w]|-|R[w \wedge v]|$.

Рассмотрим теперь те запросы, в которых явным образом содержится указание на множество соотнесения (универсум рассуждения). Эти запросы в общем случае будут иметь вид $R(u \wedge w)$. Нам необходимо определить оценки для $|R[u \wedge (w \wedge v)]|$ и $|R[u \wedge (w \vee v)]|$ на основании оценок $|R[u \wedge w]|$ и $|R[u \wedge v]|$.

Для определения интервала, в котором заключена оценка $|R[u \wedge (w \wedge v)]| = |R[(u \wedge w)] \cap R[(u \wedge w)]|$, допустим, что множества $|R[u \wedge w]|+|R[u \wedge v]| \leq |R[u]|$. В этом случае нижней границей будет 0, когда множества $R[u \wedge w]$ и $R[u \wedge v]$ не пересекаются, а верхней границей будет $\min(|R[u \wedge w]|,|R[u \wedge v]|)$, когда одно из множеств включено в другое. Теперь допустим, что $|R[u \wedge w]|+|R[u \wedge v]|>|R[u]|$. В этом случае в пересечении множеств $R[u \wedge w]$ и $R[u \wedge v]$ содержится не менее $|R[u \wedge w]|+|R[u \wedge v]|-|R[u]|$ и опять же не более $\min(|R[u \wedge w]|,|R[u \wedge v]|)$ элементов.

Приведенное рассуждение позволяет заключить, что $\max(0,|R[u \wedge w]|+|R[u \wedge v]|-|R[u]|) \leq |R[u \wedge (w \wedge v)]| \leq \min(|R[u \wedge w]|,|R[u \wedge v]|)$.

Повторим рассуждение для $|R[u \wedge (w \vee v)]|=|R[u \wedge w] \cup R[u \wedge v]|$. Если $|R[u \wedge v]|+|R[u \wedge w]|< |R[u]|$, то нижней границей для $|R[u \wedge w] \cup R[u \wedge v]|$ будет $\max(|R[u \wedge w]|,|R[u \wedge v]|)$, когда одно из множеств включено в другое, а верхней границей будет $|R[u \wedge w]|+|R[u \wedge v]|$, когда они дизъюнкты. В случае $|R[u \wedge w]|+|R[u \wedge v]| \geq |R[u]|$ нижней границей опять будет $\max(|R[u \wedge w]|,|R[u \wedge v]|)$, а верхней – $\min(|R[u]|,|R[u \wedge w]|+|R[u \wedge v]|)$.

Приведенное рассуждение позволяет заключить, что $\max(|R[u \wedge w]|,|R[u \wedge w]|) \leq |R[u \wedge (w \vee v)]| \leq \min(|R[u]|,|R[u \wedge w]|+|R[u \wedge v]|)$.

В последующем изложении будем предполагать, что у нас всегда фиксировано множество соотнесения $R[u]=U$, и потому в запросах, когда это не может привести к недоразумению, мы будем опускать u. Т.е. запросы вида $R(u \wedge w)$ станем записывать просто как $R(w)$.

Наши оценки получат более простую запись:

- $|R[-v]| = |U|-|R[v]|$
- $\max(0,|R[w]|+|R[v]|-|U|) \leq |R[w \wedge v]| \leq \min(|R[w]|,|R[v]|)$
- $\max(|R[w]|,|R[w]|) \leq |R[w \vee v]| \leq \min(|U|,|R[w]|+|R[v]|)$

Интересной представляется следующая связь полученных оценок с функциями конечнозначной логики Лукасевича [4,5].

Напомним, что матрица вида $M_L=<V_{n+1}, \sim, \rightarrow, \{n\}>$ называется n+1-значной матрицей Лукасевича ($n \in N$, $n \geq 1$), где $V_{n+1} = \{0, 1, \ldots, n-1, n\}$; $\sim$ есть унарная операция отрицания и $\rightarrow$ бинарная операция импликации, определенные на множестве V_n следующим образом:

- $\sim x = n-x$
- $x \rightarrow y = \min(n, n-x+y)$.

Операции конъюнкции и дизъюнкции вводятся по определению:

- $x \vee y = (x \rightarrow y) \rightarrow y = \max(x,y)$
- $x \wedge y = \sim(\sim x \vee \sim y) = \min(x,y)$

Определим две других операции для конъюнкции и дизъюнкции так, как мы привыкли это делать в классической логике:

- $x \# y = \sim x \rightarrow y = \min(n, x+y)$
- $x \& y = \sim(x \rightarrow \sim y) = \sim(\sim x \# \sim y) = \max(0, x+y-n)$

Если теперь посмотреть на количественные характеристики ответов на запросы, то они в $|U|+1$-значной логике Лукасевича будут выглядеть следующим образом:

- $|R[-v]| = \sim|R[v]|$

- $|R[w]| \& |R[v]| \leq |R[w \wedge v]| \leq |R[w]| \wedge |R[v]|$
- $|R[w]| \vee |R[w]| \leq |R[w \vee v]| \leq |R[w]| \# |R[v]|$

Оценка для конъюнктивного запроса ограничена сверху и снизу двумя видами конъюнкций, а оценка дизъюнктивного запроса – двумя видами дизъюнкций логики Лукасевича. Если учесть, что конъюнкции и дизъюнкции конечнозначной логики Лукасевича мы определяли только через отрицание и импликацию, то получим, что границы интервалов для количественных характеристик ответов на запросы представимы посредством одних лишь отрицания и импликации логики Лукасевича. Это дает еще один повод к осмыслению того, чем является логика Лукасевича.

Об отношении логики и теории вероятностей

Для дальнейшего изложения нам необходимо прояснить связь между логикой и теорией вероятностей. Ни для кого не является секретом так называема логическая интерпретация вероятности. Но возможен и другой концептуально отличный взгляд на их отношение, заключающийся в том, что логика и теория вероятностей являются в определенном смысле *теориями-двойниками*, изучающими одни и те же объекты, но с двух взаимодополнительных точек зрения.

Фиксируем язык логики высказываний:

1. p, q, r, s, ... – множество пропозициональных переменных;
2. &, ∨, ¬ - логические связки.

Определение формулы – обычное.

Моделью нашего языка будем называть пару M=<W, |.|>, где

1. W – множество возможных миров;
2. |.| - функция интерпретации пропозициональных переменных, сопоставляющая каждой переменной p некоторое подмножество $|p| \subseteq W$ – область ее истинности.

Обычным образом распространяем функцию |.| на множество всех формул:

1. $|\neg A| = W - |A|$

2. $|A\&B| = |A| \cap |B|$
3. $|A \vee B| = |A| \cup |B|$

Как обычно, мы говорим, что формула A *общезначима* в модели M=<W, |.|>, если имеет место |A|=W. Соответственно формула A *противоречива* в модели M=<W, |.|>, если имеет место |A|=∅.

Итак, с точки зрения логики, каждое высказывание интерпретируется некоторым множеством ситуаций/миров, в которых это высказывание истинно. Для классической логики семейство областей истинности формул образует булеву алгебру относительно операций дополнения, пересечения и объединения множеств.

Затем мы вспоминаем, что всякое множество характеризуется еще и мощностью, и решаем построить теорию количественных оценок областей истинности формул.

Для начала допустим, что мы имеем дело только с конечными моделями, в которых мощность множества возможных миров W оценивается некоторым натуральным числом N.

Мощность множества |A| будем обозначать посредством n(A). Если мы захотим определять количественные оценки сложных формул на основании количественных оценок их подформул, то обнаружим, что функциональная зависимость существует не всегда.

В случае формул вида ¬A все просто - $n(\neg A) = N - n(A)$. Для формул конъюнктивного и дизъюнктивного вида ситуация сложнее.

1. $n(A\&B) = n(A)+n(B)-n(A \vee B)$
2. $n(A \vee B) = n(A)+n(B)-n(A\&B)$

Возможна также интервальная оценка:

1. $\max(0, n(A)+n(B)-N) \le n(A\&B) \le \min(n(A), n(B))$
2. $\max(n(A), n(B)) \le n(A \vee B) \le \min(N, n(A)+n(B))$

Очевидно, что если формула A общезначима, то n(A)=N, а если формула A противоречива, то n(A)=0.

Одним из недостатков наших оценок является то, что они привязаны к конкретным значениям мощности множества W. Для того чтобы избавиться от этого недостатка, будем оценивать формулу A не в терминах количества миров n(A), в которых она истинна, а в терминах доли этих миров P(A)=n(A)/N от всего множества W. Тогда мы получим следующие соотношения:

1. $0 \le P(A) \le 1$ для любой формулы A;
2. P(A)=0, если A – противоречива;
3. $P(A \vee B) = P(A)+P(B)-P(A \& B)$.

Но это есть аксиомы классической теории вероятностей. Интервальные оценки, переписанные в виде:

1. $\max(0, P(A)+P(B)-1) \le P(A \& B) \le \min(P(A), P(B))$
2. $\max(P(A), P(B)) \le P(A \vee B) \le \min(1, P(A)+P(B))$

также известны в теории вероятностей [41,44].

Можно подумать, что ничего нового мы не получили, а всего лишь пришли к давно известной логической интерпретации теории вероятностей. Да, подумать можно, но все-таки это не так. Важно не только то, что мы получили в результате, но и мотивация наших действий, которая привела к конечному результату. Дело в том, что мы не ставили себе цели дать логическую интерпретацию теории вероятностей. Нас интересовала *теория количественных оценок областей истинности формул*, развив которую, мы пришли к выводу что для классической логики эта теория в точности совпадает с теорией вероятностей.

При переходе от логики высказываний к логике предикатов мы можем интересоваться количественными оценками множества приписываний в фиксированной модели, которые выполняют данную формулу, или количественными оценками множества моделей, в которых истинна данная формула. Переход от конечных моделей к бесконечным может быть произведен просто как обобщение конечного случая.

Мы уже давно отказались от мысли о существовании единственно истинной логики. Одним из важных следствий полученного нами результата будет отказ от единственно истинной теории вероятностей. Каждой неклассической логике будет соответствовать ее теория количественных оценок для областей истинности формул. Эти теории в общем случае не будут совпадать с теорией количественных оценок для классической логики. Например, если в классической теории имеет место $P(A \vee \neg A) = P(A)+P(\neg A) = 1$, то очевидно в интуиционистской логике это выполняться не будет, так как в ней формула $A \vee \neg A$ не является теоремой [45]. Для паранепротиворечивых логик не будет выполняться соотношение $P(A \& \neg A) = 0$. Очень интересными являются теории количественных оценок областей истинности для многозначных логик.

Логики-философы давно и глубоко занимаются вопросом, как влияет принятие тех или иных онтологических и гносеологических предпосылок на принятие различных законов логики. Теперь можно точно так же поставить вопрос об аналогичном влиянии онтологических и гносеологических предпосылок на количественные и вероятностные оценки предметной области. Это имеет важное прикладное значение, так как мы являемся свидетелями экспансии неклассических логик в науке и технике, а переход от качественных к количественным оценкам расширяет сферу их применимости. Далеко не случайно появилось много работ по квантитативной силлогистике, в которой терминам сопоставляются не множества, как мы привыкли, а именно оценки числа входящих в них элементов [37,38,39,40]. Но силлогистика представляет собой лишь небольшой фрагмент науки логики.

Теперь посмотрим, как соотносится с логикой статистика.

В колмогоровской модели теории вероятностей имеется семейство подмножеств множества V, замкнутое относительно объединения, пересечения и дополнения, и функция P, определенная на этом семействе со значениями в замкнутом интервале от 0 до 1.

Функция P удовлетворяет трем аксиомам:
1. $0 \leq P(A) \leq 1$

2. $P(\varnothing) = 0$
3. $P(A \cup B) = P(A)+P(B)-P(A \cap B)$

В общем виде задачей статистики является обратная задача восстановление функции P на основании некоторой ограниченной выборки элементов множества V. В привычной нам науке логики эту же задачу, но другими средствами, призвана решать индуктивная логика. Т.е. мы получаем, что *количественным двойником дедуктивной логики является теория вероятностей, а количественным двойником индуктивной логики является теория статистики.* Подводя итог, можно сказать, что допустим взгляд на теорию вероятностей как на *квантитативную логику.*

Если пойти немного дальше и выйти за рамки отношения логики и теории вероятностей, то можно вспомнить, что множества, помимо оценки их мощности, могут характеризоваться типом упорядочения. Т.е. можно развить теорию порядковых оценок областей истинности. При этом порядок не обязан быть линейным, а может иметь гораздо более сложную структуру. Замечу, что некоторые работы логиков имеют к этому отношение. Достаточно привести пример *алгебр как истинностных значений* из работы [4].

Вероятностная модель запросов

Причина, почему так важно уметь оценивать количественные характеристики ответов на запросы, кроется в том, что множество, представляющее объем ответа, дано пользователям Интернет всегда потенциально, а количественная характеристика дана актуально. Например, в момент написания этих слов в поисковой системе AltaVista был сделан запрос *найти все страницы, на которых упоминаются слова United States*. В ответ были выданы первые десять адресов страниц, удовлетворяющих нашему запросу, и сообщение, что всего таких страниц около 959.000.000 (девятьсот пятьдесят девять миллионов). Понятно, что никто и никогда не сможет просмотреть все эти страницы, но количественная характеристика ответа может быть использована. Последующее изложение как раз и будет связано с тем, как можно использовать количественные характеристики для извлечения из сети Интернет дополнительной информации.

Пусть дано множество соотнесения U, состоящее из |U| элементов. Сопоставим каждому подмножеству S множества U число P(S), заключенное в интервале $0 \le P(S) \le 1$ и представляющее вероятность того, что случайно выбранный элемент $u \in U$ будет принадлежать множеству S. Так как выбор элемента u случаен, т.е. равновероятен для всех элементов множества U, определение P(S) будет выглядеть следующим образом:

Def.19 $P(S) = |S|/|U|$, где $S \in U$

Т.е. P(S) сопоставлено число, представляющее ту долю, которую составляют элементы множества S от числа всех элементов множества U. В согласии с традицией теории вероятностей будем называть подмножества множества U событиями.

Семейство подмножеств множества U замкнуто относительно дополнения, объединения и пересечения, т.е. является алгеброй множеств. Очевидно, что для этих множеств будут выполняться следующие условия, являющиеся аксиомами теории вероятностей:

P 1. $P(\varnothing) = 0$;
P 2. $0 \le P(S) \le 1$ для любого $S \in U$;
P 3. $P(S \cup V) = P(S) + P(V)$, если и только если $S \cap V = \varnothing$.

Таким образом, на множестве адресов Интернет-страниц, удовлетворяющих условиям произвольного запроса u, мы всегда можем естественным образом задать дискретное вероятностное пространство.

Как и положено, для двух событий $S \subseteq U$ и $V \subseteq U$ определим условную вероятность того, что если произвольный элемент $u \in U$ принадлежит множеству V, то он также будет принадлежать множеству S.

Def.20 $P(S/V) = P(S \cap V)/P(V)$ ($= |S \cap V|/|V|$)

Смысл именно такого определения очевиден. Так как мы интересуемся лишь элементами множества V, то дробь $P(S \cap V)/P(V)$ в определении как раз и представляет, какую долю от множества V составляют элементы множества S.

Ниже приводим некоторые теоремы теории вероятностей.

P 4. $P(S \cup V) = P(S) + P(V) - P(S \cap V)$
P 5. $P(S/S \cap V) = 1$
P 6. $P(S/V) + P(-S/V) = 1$
P 7. $P(-S/S \cap V) = 0$
P 8. $P(S \cup V/W) = P(S/W) + P(V/W) - P(S \cap V/W)$
P 9. $S \subseteq W \Rightarrow P(S) \leq P(W)$

Необходимо отметить, что мы не можем определить вероятностного пространства на множестве всех адресов Интернет-страниц, так как для пользователя оно не является алгеброй множеств.

Если R(w) – запрос к сети Интернет, то вместо P(R[w]) будем писать просто P(w) и читать как *вероятность того, что случайно выбранный из множества соотнесения адрес удовлетворяет запросу R(w)*, т.е. принадлежит его области истинности R[w]. Соответственно запись P(w/v) на самом деле будет служить сокращением для P(R[w]/R[v]).

Так как $R[w \wedge v] = R[w] \cap R[v]$, то $P(w \wedge v) = P(R[w] \cap R[v])$. Аналогично $P(w \vee v) = P(R[w] \cup R[v])$ и $P(-w) = P(R[-w]) = P(U \backslash R[w]) = 1 - P(R[w])$.

Всякий ответ на запрос к поисковой системе является событием в построенном нами вероятностном пространстве.

Дадим определение понятия независимых событий, которое будет являться центральным в последующем изложении.

Def.21 События S и V *независимы*, если и только если $P(S \cap V) = P(S)*P(V)$.

Смысл этого определения становится ясен, если равенство преобразовать к следующему виду $P(S \cap V)/P(V) = P(S)$. Слева от знака равенства $P(S \cap V)/P(V)$ представляет условную вероятность P(S/V), т.е. долю, которую составляют элементы множества S от множества V, а справа стоит P(S) – оценка доли, которую составляют элементы множества S от всего множества соотнесения U. Получаем, что события S и V независимы, если и

только если эти доли равны. Не количества, а именно доли, которые всегда выражаются дробными числами в интервале от 0 до 1.

Дадим применительно к нашей модели еще одно истолкование независимости событий, но уже в терминах количества элементов. Для этого в соответствии с определением $P(S \cap V)$ и $P(V)$ перепишем равенство $P(S \cap V) = P(S)*P(V)$ следующим образом $|S \cap V|/|U| = P(S)*|V|/|U|$. Умножив обе части на $|U|$, получим $|S \cap V| = P(S)*|V|$. Т.е. события S и V независимы, если и только если количество элементов множества $S \cap V$ равно P(S), умноженному на количество элементов множества V.

Подтверждение и принятие гипотез

Элементы теории вероятностей понадобились нам для того, чтобы с ее помощью определять, когда ответы на запросы в построенном нами вероятностном пространстве являются зависимыми событиями.

Вывод о том, что два события являются зависимыми, производится по правилу, напоминающему *modus tollendo tollens* $A \supset B, \neg B \Rightarrow \neg A$.

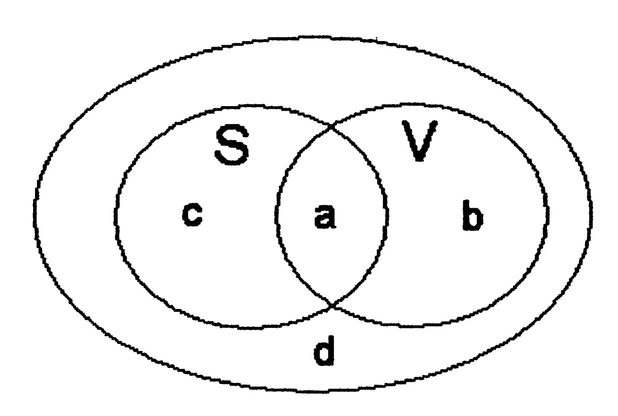
Пусть даны два события S, V. Из гипотезы, что эти события независимы H(S,V), в теории статистики следует, что с большой вероятностью они должны находиться в определенном отношении $\varphi(S,V)$, которое может быть проверено эмпирически. Посредством обращения к свойствам построенного вероятностного пространства мы показываем, что имеет место $\neg\varphi(S,V)$, вероятность истинности которого в силу одних лишь случайных причин не превышает некоторого достаточно малого значения α. Отсюда делается вывод, что на уровне значимости α имеет место $\neg H(S,V)$.

Уточнение правила опровержения гипотез будет выглядеть следующим образом:

$$H(S,V) \supset P(\varphi(S,V)) \geq 1-\alpha,\ \neg\varphi(S,V)\ \Rightarrow \neg H(S,V)_{\alpha}$$

Для практического использования данного правила нам осталось лишь уточнить, что из себя представляет $\varphi(S,V)$.

Есть много конкретных вариантов того, что использовать в качестве $\varphi(S,V)$. В их основе лежит сопоставление ожидаемой вероятности $Pe(S \cap V) = P(S)*P(V)$ при независимости событий S и V с фактической $Pf(S \cap V)$. Мы будем использовать так называемый непараметрический критерий Фишера [1,9].



Построим следующую таблицу (сопряженности):

	S	-S
V	a	b
-V	c	d

где $a=|S \cap V|$, $b=|-S \cap V|$, $c=|S \cap -V|$, $d=|-S \cap -V|$, т.е. количества элементов, вошедших в соответствующие пересечения. При независимости S и V мы ожидаем, например, что величина $Pe(S \cap V)$ будет равна $P(S)*P(V) = |S|*|V|/(|U|*|U|) = (a+c)*(a+b)/((a+b+c+d)*(a+b+c+d))$.

Критерий Пирсона-Фишера позволяет сравнивать ожидаемые при условии независимости частоты с фактическими и на основании такого сравнения делать выводы о независимости/зависимости событий. Теорема Пирсона-Фишера гласит, что при неограниченном росте $|U|=a+b+c+d$ случайная величина, рассчитываемая по формуле

Def.22 $X2 = (a+b+c+d)*(a*d-b*c)*(a*d-b*c)/((a+c)*(a+b)*(b+d)*(c+d))$,

стремится к распределению *хи-квадрат*. В случае зависимых признаков данная величина неограниченно возрастает.

Имеются специально рассчитанные таблицы критических значений X2, при превышении которых можно с различной степенью вероятности утверждать, что события S и V не являются независимыми. Фрагмент такой таблицы приводится ниже.

$\alpha=0,1$	$\alpha=0,05$	$\alpha=0,01$	$\alpha=0,001$
2,71	3,84	6,63	10,83

Например, если рассчитанная величина X2 превысила 6,63, то вероятность того, что данное событие случайно, не превышает 0,01. Следовательно, с вероятностью не менее 0,99 мы можем утверждать, что события S и V зависимы.

Таким образом, в качестве $\varphi(S,V)$ мы рассматриваем утверждение «*величина X2 не превышает критических значений*».

После прихода к заключению о наличии связи между двумя событиями необходимо исследовать сам характер этой связи. Дело в том, что мы пока не знаем направленности связи. Если $P(S\cap V)>P(S)*P(V)$, то связь событий S и V *положительна*, так как их пересечение содержит больше элементов, чем ожидалось бы в случае их независимости. Аналогично, если $P(S\cap V)<P(S)*P(V)$, то связь событий S и V *отрицательна*, так как их пересечение содержит меньше элементов, чем ожидалось бы в случае их независимости.

Def.23 $R(w)\approx_\alpha R(v)$ – запрос R(w) *ассоциативно связан с запросом* R(v) на уровне α, если и только если гипотеза о независимости событий R[w] и R[v] отвергнута на уровне α и $P(v\wedge w)/(P(v)*P(w))>1$.

Def.24 $R(w)\Diamond_\alpha R(v)$ – запрос R(w) *диссоциативно связан с запросом* R(v) на уровне α, если и только если гипотеза о независимости событий R[w] и R[v] отвергнута на уровне α и $P(v\wedge w)/(P(v)*P(w))<1$.

Заметим, что $P(v\wedge w)/(P(v)*P(w)) = P(v/w)/P(v) = a*d/(b*c)$. Эта оценка может принимать любые неотрицательные значения и

показывает, во сколько раз больше вероятность встретить страницу, которая одновременно удовлетворяет двум запросам R(v) и R(w), по сравнению с ожидаемой вероятностью, если бы запросы были независимы друг от друга. Значения меньше единицы соответствуют отрицательной связи между событиями, а значения больше единицы – положительной связи.

Интересно вновь обратиться к конечнозначной логике Лукасевича. Можно показать, что определяемые в ней конъюнкция & и дизъюнкция #, задают диссоциативную связь между высказываниями, а конъюнкция $\wedge$ и дизъюнкция $\vee$ задают ассоциативную связь. Представляет интерес расширение логики Лукасевича новым видом конъюнкции, соответствующим независимым высказываниям. На уровне семантики это просто произведение значений конъюнктов. При этом, разумеется, придется совершить переход от конечнозначной к бесконечнозначной логике. Такое расширение является переходом к product-логике Лукасевича. В последние годы исследователи многозначных и нечетких логик уделяют много внимания именно product-логикам.

Справедливы следующие правила вывода:

R 1. $R(w) \approx_\alpha R(v) \Rightarrow R(v) \approx_\alpha R(w)$
R 2. $R(w) \approx_\alpha R(v) \Rightarrow R(-w) \approx_\alpha R(-v)$
R 3. $R(w) \approx_\alpha R(v) \Rightarrow R(w) <>_\alpha R(-v)$
R 4. $R(w) <>_\alpha R(v) \Rightarrow R(w) \approx_\alpha R(-v)$

Покажем, что эти правила действительно имеют место.

Правило R.1 очевидно и его доказательство мы опускаем.

Доказательство R2.
Допустим, что $R(w) \approx_\alpha R(v)$. По Def.23 гипотеза о независимости R[w] и R[v] отвергнута на уровне α и $P(v \wedge w)/(P(v)*P(w)) > 1$. Это означает, что величина X2, вычисляемая по правилу Def.22 превысила некоторое пороговое значение для уровня значимости α.

Таблица сопряженности для R[w] и R[v] выглядит следующим образом:

	R[w]	-R[w]
R[v]	a	b
- R[v]	c	d

X2 = (a+b+c+d)*(a*d-b*c)*(a*d-b*c)/((a+c)*(a+b)*(b+d)*(c+d))

и

$P(v\wedge w)/(P(v)*P(w)) = P(v/w)/P(v) = a*d/(b*c) > 1$.

Построим таблицу сопряженности для -R[w] и -R[v]. Она будет выглядеть следующим образом:

	-R[w]	R[w]
-R[v]	d	c
R[v]	b	a

X2 = (a+b+c+d)*(d*a-c*b)*(d*a-c*b)/((b+d)*(d+c)*(c+a)*(b+a)) = (a+b+c+d)*(a*d-b*c)*(a*d-b*c)/((a+c)*(a+b)*(b+d)*(c+d)) и

$P(-v\wedge -w)/(P(-v)*P(-w)) = a*d/(b*c)>1$.

Таким образом, гипотеза о независимости -R[w] и -R[v] также может быть отвергнута на уровне α и следовательно $R(-w)\approx_{\alpha}R(-v)$.

Доказательство R.3

Допустим, что $R(w)\approx_{\alpha}R(v)$

По Def.23 гипотеза о независимости R[w] и R[v] отвергнута на уровне α и $P(v\wedge w)/(P(v)*P(w))>1$. Это означает, что величина X2, вычисляемая по правилу Def.22 превысила некоторое пороговое значение для уровня значимости α.

Таблица сопряженности для R[w] и R[v] выглядит следующим образом:

	R[w]	-R[w]
R[v]	a	b
- R[v]	c	d

X2 = (a+b+c+d)*(a*d-b*c)*(a*d-b*c)/((a+c)*(a+b)*(b+d)*(c+d))

и

$P(v\wedge w)/(P(v)*P(w)) = P(v/w)/P(v) = a*d/(b*c) > 1$.

Построим таблицу сопряженности для R[w] и -R[v]. Она будет выглядеть следующим образом:

	R[w]	-R[w]
-R[v]	c	d
R[v]	a	b

$X2 = (a+b+c+d)*(c*b-a*d)*(c*b-a*d)/((c+a)*(c+d)*(d+b)*(a+b))$
$= (a+b+c+d)*(a*d-b*c)*(a*d-b*c)/((a+c)*(a+b)*(b+d)*(c+d))$

$P(-v \wedge w)/(P(-v)*P(w)) = c*b/(a*d)<1$

Таким образом, гипотеза о независимости R[w] и -R[v] также может быть отвергнута на уровне α и следовательно $R(w) <>_{\alpha} R(-v)$.

Доказательство R.4 затруднений не вызывает.

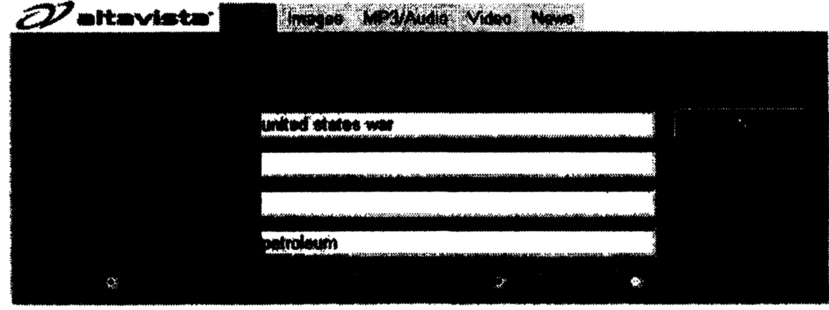
Практический пример 1

Приведем пример того, как все это может работать на практике. Рассмотрим следующее утверждение. *В наше время скрытой причиной многих войн является борьба за контроль над нефтяными ресурсами.* Оно принимается нами как верное, но в то же время трудно припомнить хотя бы одну войну, целью которой явно декларировался контроль за нефтяными ресурсами.

Нас будет интересовать вопрос, действительно ли в наше время имеется связь между нефтью и войнами, которые ведутся в мире. При запросе к сети Интернет в качестве множества соотнесения возьмем лишь те страницы, в которых содержится упоминание *United States*.

Итак, пусть u=*United States*, w=*war*, p=*petroleum*. Нам необходимо получить количественные оценки ответов на запросы $u \wedge w \wedge p$, $u \wedge w \wedge -p$, $u \wedge -w \wedge p$ и $u \wedge -w \wedge -p$.

Воспользуемся поисковой системой Интернет AltaVista. Запрос $u \wedge w \wedge -p$ и ответ на него будет выглядеть в ней следующим образом



В сети Интернет на момент написания настоящей работы было около 137 миллионов страниц, на которых в одном контексте упоминались *United States* и *war*, и не упоминалось *petroleum*.

Выполним остальные три запроса и составим таблицу сопряженности:

United States	petroleum	- petroleum
war	3,98 млн.	137 млн.
- war	7,47 млн.	806 млн.

Вычисляем оценку X2 = 3678214. Сравниваем ее с таблицей критических значений и видим, что мы можем на уровне 0,001 отвергнуть гипотезу о независимости событий R[u∧w] и R[u∧p].

1	P(u∧w)	0,14771
2	P(u∧p)	0,01199
3	P(u∧w)*P(u∧p)	0,00177

4	$P(u \wedge w \wedge p)$	0,00417
5	$P(u \wedge war/u \wedge p)$	0,34760
6	$P(u \wedge war/u \wedge \neg p)$	0,14528
7	$P(u \wedge p/u \wedge w)$	0,02823
8	$P(u \wedge w \wedge p)/(P(u \wedge w)*P(u \wedge p))$	2,35593

Так как $P(u \wedge w \wedge p)/(P(u \wedge w)*P(u \wedge p)) = 2,35593 > 1$, мы получаем, что на множестве соотнесения *United States* запрос R(war) ассоциативно связан с запросом R(petroleum) на уровне 0,001. В нашей записи $R(u \wedge w) \approx_{0,001} R(u \wedge p)$.

Из $P(u \wedge war/u \wedge p) = 0,34760$ следует, что на каждой третьей Интернет-странице, на которой упоминается *petroleum*, упоминается и *war*. В то же время из $P(u \wedge war/u \wedge \neg p) = 0,14528$ следует, что лишь на каждой седьмой Интернет-странице, на которой не упоминается *petroleum*, содержится упоминание *war*.

Полученный результат конечно же не является доказательством того, что *в наше время скрытой причиной многих войн является борьба за контроль над нефтяными ресурсами*. В то же время нельзя отрицать очевидного факта, что содержание Интернет-публикаций далеко не случайно, а отражает происходящие в реальной жизни события, ожидания людей и пр. Поэтому сильная ассоциативная связь между страницами, содержащими слово *petroleum*, и страницами, содержащими слово *war*, говорит о существовании реальной связи между событиями, имеющими отношение к войне и нефти. Интересно то, что эта связь не обязательно должна быть явно осознана авторами публикаций. От них лишь требуется быть добросовестными регистраторами всего, что происходит вокруг. Это мы уже будем искать скрытые закономерности в том, что они увидели и зафиксировали.

Ряды событий

Закономерности, которые можно искать в сети Интернет, не ограничиваются одними лишь ассоциативными связями между отдельными событиями. Значительный интерес представляет поиск связей между рядами событий.

Пусть имеются две функции f и g, определенные на одной области I. Допустим, для конечного множества элементов из

области определения $\{i_1,\dots,i_n\}\subseteq I$ нам известны значения функций $f(i_1)=x_1,\dots,f(i_n)=x_n$, $g(i_1)=y_1,\dots,g(i_n)=y_n$. Требуется ответить на вопрос, имеется ли какая-нибудь связь между функциями f и g?

Следует обратить внимание не то, что данная постановка задачи имеет много общих черт с задачами, для решения которых предложил свои методы Д.С.Милль. Фактически речь пойдет об одном из современных уточнений метода сопутствующих изменений.

Существующие закономерности можно разделить на детерминистические и недетерминистические. При этом детерминистические закономерности являются всего лишь предельным случаем недетерминистических.

Например, пусть I – это множество людей, f сопоставляет каждому человеку его рост, а g - его вес. Очевидно, что по значению одного параметра нельзя предсказать точное значение другого, но в то же время очевидно, что связь между ростом и весом человека имеется. Обычно, чем больше рост, тем больше вес, и чем больше вес, тем больше рост. Для оценивания такого рода закономерностей, которые могут связывать значения параметров лишь приблизительно, используются специальные оценки под названием *коэффициент корреляции*.

Для тех случаев, когда функции f и g принимают числовые значения, может использоваться коэффициент линейной корреляции Пирсона, вычисляемый по формуле:

Def.25 $r = \sum_i(x_i-X)*(y_i-Y)/(\sqrt{\sum_i(x_i-X)^2}\sqrt{\sum_i(y_i-Y)^2})$, где $X=\sum_i x_i/n$, $Y=\sum_i y_i/n$.

Коэффициент корреляции может принимать значения от -1 до +1. Если он отрицателен, то чем больше (меньше) значение одного параметра, тем меньше (больше) значение другого параметра. Если же коэффициент корреляции положителен, то чем больше (меньше) значение одного параметра, тем больше (меньше) значение другого параметра. Крайние значения -1 или +1 указывают на то, что по значению одного параметра мы можем с абсолютной точностью предсказать значение другого параметра. В случае роста и веса людей коэффициент корреляции

положителен, но все-таки меньше единицы, так как связь между параметрами не является однозначной.

Специальные формулы для вычисления коэффициентов корреляции предложены для тех случаев, когда область значений функций f и g не является числовой, а всего лишь линейно упорядочена. Это так называемые коэффициенты ранговой корреляции Спирмена и Кенделла.

В предельном случае функции f и g могут быть просто булевозначными. Если считать, что они принимают значения 1 и 0, то формула для вычисления коэффициента корреляции остается той же, что и для линейной корреляции Пирсона, но может быть выражена несколько проще:

Def.26 $r=(p_{ij}-p_i*p_j)/\sqrt{(p_i*(1-p_i)*p_j*(1-p_j))}$, где $p_{ij}=\sum x_i*y_i/n$, $p_i=\sum x_i/n$, $p_j=\sum y_i/n$

Если область определения функций f и g линейно упорядочена, например, в случае, когда она является множеством моментов времени, открываются дополнительные возможности для анализа. Пусть нас интересует связь между государственным финансированием образования и темпами развития экономики. Если в качестве анализируемых данных мы возьмем соответствующие величины для разных стран, то получим коэффициент корреляции, связывающий эти два параметра.

Аналогичное исследование можно провести и для отдельно взятой страны, взяв в качестве анализируемых данных ежегодные вложения в образование и ежегодные оценки темпов развития экономики. Во втором случае мы будем иметь дело с анализом временных рядов. Очевидно также, что во втором случае может быть поставлена задача выявления причинной зависимости между анализируемыми параметрами. Особенность здесь заключается в том, что увеличение или уменьшение финансирования образовательной сферы сказывается на темпах развития экономики не сразу, а с задержкой на несколько лет. Знание таких закономерностей позволяет заблаговременно прогнозировать наступление негативных событий и принимать меры для их предотвращения.

Подтвердить гипотезу о существовании причинной зависимости между двумя временными рядами данных можно путем вычисления так называемой кросскорреляции этих рядов. Это набор коэффициентов корреляции для различных временных сдвигов двух рядов друг относительно друга. Например, при вычислении коэффициентов корреляции мы сравниваем финансирование образования с темпами экономического развития через год, через два, через три и т.д. Или наоборот сравниваем финансирование образования с темпами экономического развития, какими они были год назад, два года назад, три и т.д. В зависимости от того, в какую сторону осуществлен временной сдвиг, проверяются два варианта гипотезы о направленности причинной связи.

Вывод о наличии зависимости между двумя функциями f и g производится по правилам, аналогичным тем, которые мы использовали для выяснения зависимости между отдельными событиями.

Пусть даны две функции f, g. Из гипотезы, что они независимы H(f,g), в теории следует, что с большой вероятностью они должны находиться в определенном эмпирически проверяемом отношении λ(f,g), которое выражается с помощью коэффициента корреляции. Посредством обращения к свойствам коэффициента корреляции мы показываем, что имеет место ¬λ(f,g), вероятность истинности которого для конкретных эмпирических данных не превышает некоторого достаточно малого значения α. Отсюда делается вывод, что на уровне значимости α имеет место ¬H(f,g).

$$H(f,g) \supset P(\lambda(f,g)) \geq \alpha,\ \neg\lambda(f,g)\ \Rightarrow \neg H(f,g)_{\alpha}$$

Для практического использования данного правила нам осталось лишь уточнить, что из себя представляет λ(f,g).

В случае коэффициента линейной корреляции Пирсона по формуле abs(r)*√(n-1) вычисляется оценка и для выбранного уровня значимости α сравнивается с критическими значениями специальных таблиц. Если оценка превосходит табличное значение, то вероятность данного события для независимых функций не превышает уровня значимости α. Например, для

случая n=10 таблица критических значений имеет следующий вид:

α=0,1	α=0,05	α=0,01	α=0,001
1,65	1,9	2,29	2,62

Практический пример 2

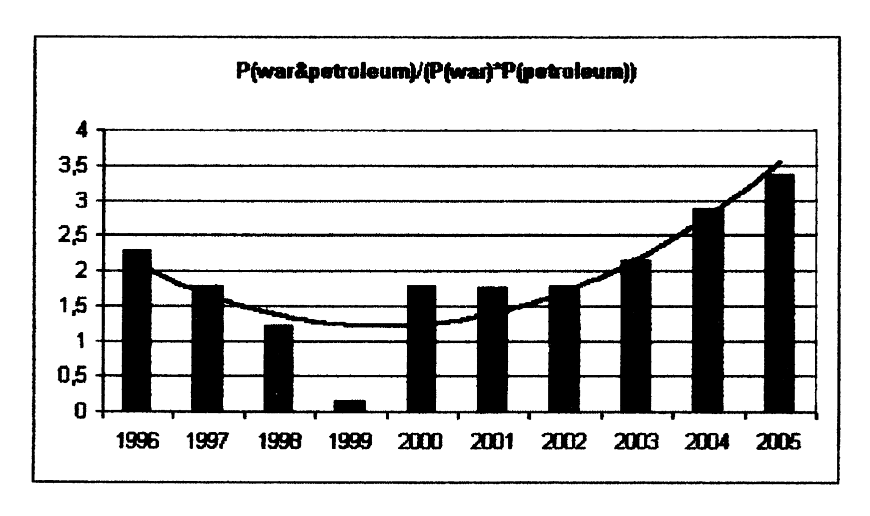
Покажем, как это будет работать в применении к Интернет-запросам.

Мы уже знаем, что запрос R(war) ассоциативно связан с запросом R(petroleum). Проведем такое же исследование, но отдельно по последним десяти годам. Современные поисковые системы, например, AltaVista и Яndex, позволяют указывать временной интервал, которому должны принадлежать страницы из области истинности запроса. Результаты приведены в следующей таблице. Из предпоследнего столбца видно, что оценка X2 в каждой строке превышает критические значения для α=0,001. Раз так, то особый интерес для нас будет представлять последний столбец, в котором стоит оценка силы ассоциативной связи между запросами.

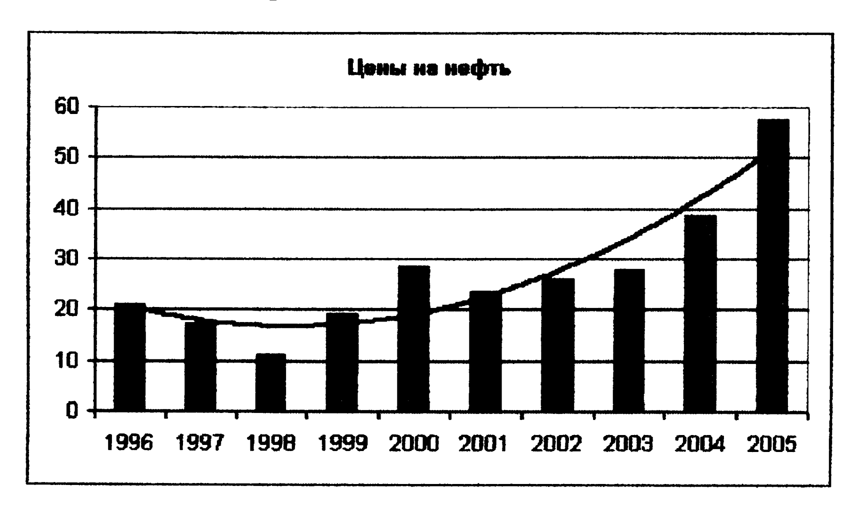
Date	w∧p	¬w∧p	w∧¬p	¬w∧¬p	X2	P(w∧p)/(P(w)*P(p))
1996	389	782	75700	349129	188	2,2942085
1997	763	1510	210000	738727	171	1,7775109
1998	1120	3220	367000	1278660	30	1,2118564
1999	2020	48600	609000	2130380	9668	0,1453968
2000	37200	70700	981000	3311100	7991	1,775934
2001	66300	114000	1740000	5279700	13434	1,764691
2002	92600	164000	2560000	8083400	19708	1,7828764
2003	178000	239000	4580000	13203000	60490	2,1469862
2004	376000	411000	8050000	25363000	232285	2,8823769
2005	2870000	5270000	110000000	679860000	3019029	3,3658758

Стоит обратить внимание на то, что в 1999 оценка ассоциативной связи приняла значение меньше 1, т.е. в 1999 году связь между R(war) и R(petroleum) была диссоциативной.

Для наглядности представим эти же оценки в виде диаграммы. Черная кривая линия – это аппроксимация значений полиномом 2-й степени.



Следующая диаграмма представляет августовские цены на нефть за этот же период.



Очень похожие диаграммы.

Допустим, что оценка ассоциативной связи между R(war) и R(petroleum) и *цены на нефть* не зависят друг от друга. Вычислим коэффициент линейной корреляции между ними r=0,7799. Величина abs(r)*√(n-1) = 0,7799*√9 = 2,3399 превосходит критическое значение 2,29 для α=0,01. Т.е. это дает нам основания отвергнуть гипотезу о независимости и с вероятностью больше 0,99 утверждать о существовании положительной зависимости/корреляции между двумя рядами значений.

Вычислим еще два коэффициента корреляции. Первый r1=0,4294 – между значениями P(w∧p)/(P(w)*P(p) за 1996-2004 гг. и *ценами на нефть* за 1997-2005 гг. Второй r2=0,9208 - между значениями P(w∧p)/(P(w)*P(p) за 1997-2005 гг. и *ценами на нефть* за 1996-2004 гг.

Коэффициент корреляции r1=0,4294 между P(w∧p)/(P(w)*P(p) и ценой на нефть в следующем году является положительным, но не дотягивает до значимых уровней. Поэтому мы не можем со сколько-нибудь большой вероятностью утверждать о временной связи между двумя рядами значений. А вот коэффициент корреляции r2=0,9208 между сегодняшней ценой на нефть и значениями P(w∧p)/(P(w)*P(p) в следующем году достаточно велик, чтобы с вероятностью больше 0,999 утверждать о существовании положительной временной связи. В данной ситуации интересно то, что мы не можем утверждать, что сегодняшняя цена на нефть положительно или отрицательно зависит от прошлогоднего P(w∧p)/(P(w)*P(p), но зато мы можем утверждать, что сегодняшняя цена на нефть влияет на то, как сильно в следующем году будет ассоциироваться R(war) и R(petroleum). Чем выше сегодняшняя цена на нефть, тем чаще в следующем году люди будут в одном контексте упоминать *war* и *petroleum*.

Приведенный выше пример интересен тем, что он устанавливает связь между моделью Интернета Mi и моделью внешнего мира Mw, показывает неслучайность информации, публикуемой на страницах глобальной сети. Как уже было сказано выше, корреляционные связи являются одним из современных уточнений методов индуктивной логики Д.С.Милля, которые направлены на обнаружение в том числе и

причинных связей. При умелом подходе эта информация может быть использована в прогностических целях, чего ожидать в ближайшем будущем в реальном мире. Мы еще вернемся к таким примерам чуть позже.

Практический пример 3

Ряды событий имеет смысл анализировать не только с целью последующего вычисления коэффициентов корреляции.

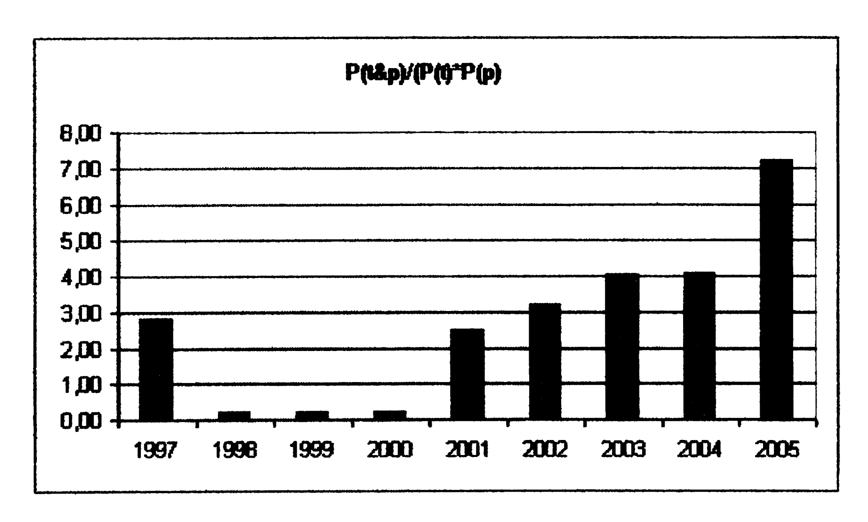
Возьмем в качестве множества соотнесения страницы из области истинности запроса R(United States). Нас будет интересовать связь между R(terrorism) и R(poverty) во временном интервале с1997 по 2005 гг.

t – terrorism
p – poverty

Ниже приведена таблица оценок, полученных с помощью поисковой системы AltaVista. Последние два столбца – это оценка X2 и оценка силы ассоциативной связи. Последняя строка – суммарная оценка всего анализируемого периода.

Date	t∧p	-t∧p	t∧-p	-t∧-p	X2	P(t∧p)/(P(t)*P(p))
1997	255	56945	1425	898375	254	2,82
1998	754	107246	49746	1502254	2152	0,21
1999	991	169009	74809	2545191	3119	0,20
2000	2370	303630	133630	4030370	5728	0,24
2001	61800	405200	392200	6450800	42238	2,51
2002	128000	551000	693000	9628000	135876	3,23
2003	218000	812000	1082000	16288000	330354	4,04
2004	411000	1429000	2139000	30321000	627465	4,08
2005	4710000	16490000	29090000	734710000	16964639	7,21
All	5533295	20325705	33656214	806801786	17572484	6,53

Как видим, имеет место сильная ассоциативная связь между R(terrorism) и R(poverty). Но она не была постоянной, а с течением времени менялась. Лучше всего это видно на следующей диаграмме.



В 1998-2000 гг. связь между R(terrorism) и R(poverty) была даже диссоциативной, и лишь с 2001 года (террористическая атака на башни-близнецы в США) связь между R(terrorism) и R(poverty) стала ассоциативной и продолжает усиливаться. Это означает, что после 2001 года люди все чаще усматривают связь между такими явлениями как терроризм и бедность.

Было бы интересно вычислить корреляцию с числом террористических актов по годам, но мы в настоящий момент такими данными не располагаем.

Практический пример 4

Еще один пример анализа динамики ассоциативных связей относится уже к России.

Возьмем в качестве множества соотнесения страницы из области истинности запроса R(Russia). Нас будет интересовать связь между R(Putin), R(bad) и R(good) во временном интервале с 2000 по 2005 гг.

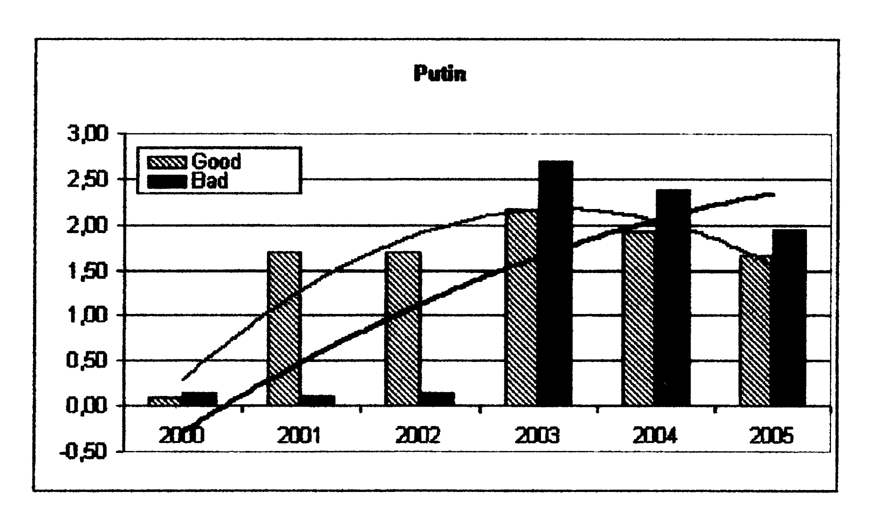
p – Putin
g – good
b – bad

Ниже приведены две таблицы оценок, полученных с помощью поисковой системы AltaVista. Последние два столбца – это оценка X2 и оценка силы ассоциативной связи.

R(Putin) - R(good)						
Date	p∧g	p∧-g	-p∧g	-p∧-g	X2	P(p∧g)/(P(p)*P(g)
2000	2490	43310	378510	605690	20475	0,09
2001	39200	43800	587800	1119200	5693	1,70
2002	47800	50400	790200	1421600	6821	1,71
2003	85200	68800	1284800	2241200	22523	2,16
2004	152000	145000	2328000	4265000	31062	1,92
2005	2520000	4350000	40680000	1,16E+08	395093	1,66

R(Putin) - R(bad)						
Date	p∧b	p∧-b	-p∧b	-p∧-b	X2	P(p∧b)/(P(p)*P(b)
2000	1270	44530	168730	815470	6559	0,14
2001	1650	81350	265350	1441650	11462	0,11
2002	2580	95620	355420	1856380	12973	0,14
2003	51800	102200	560200	2965800	33526	2,68
2004	90600	206400	1029400	5563600	46296	2,37
2005	1350000	5520000	17550000	1,4E+08	464389	1,95

На следующей диаграмме приведена оценка ассоциативной связи между R(Putin), R(good) и R(bad).



Мы видим, что в первый год первого президентского срока Путина имела место сильная диссоциативная связь между R(Putin), R(good) и R(bad). Это можно проинтерпретировать как то, что к нему присматривались и старались не выносить оценочных суждений. В 2001-2002 гг. деятельность Путина в основном оценивали положительно. В 2003 году произошел резкий всплеск негативных оценок. При более детальном анализе можно показать, что это произошло после ареста Ходорковского. С тех пор негативные оценки превалируют. В 2003 году Путин совершил главную ошибку, в результате которой отношение к нему западного сообщества резко изменилось.

Очевидно, что в данном случае вычисление корреляционных связей вряд ли возможно, но это вовсе не означает, что при анализе рядов событий мы должны ограничиваться лишь ими.

МАТЕМАТИЧЕСКИЕ МЕТОДЫ КОНТЕНТ-АНАЛИЗА

В предыдущей главе было показано, какие скрытые возможности для выявления закономерностей имеются, но не используются в уже существующих поисковых системах Интернет. Теперь же речь пойдет о расширении числа методов обработки информации, хранящейся в глобальной сети. Основная идея заключается в том, чтобы снабдить ответы на запросы некоторыми дополнительными оценками, базирующимися на частотных характеристиках встречаемости слов на страницах Интернет. Эта информация никак не представлена в стандартных ответах на запросы, но она хранится в базах данных поисковых систем и используется для оценки степени релевантности Интернет-страниц конкретным запросам. Таким образом, мы приближаемся к возможности совмещения поиска информации в сети Интернет с контент-анализом.

Что такое контент-анализ?

Контент-анализ официально существует уже более ста лет, но до сих пор имеется целый ряд заблуждений относительно того, что же он из себя представляет. Очень часто этот термин дословно переводят на русский язык как "анализ содержания" и считают, что все поняли, что это просто содержательный анализ текстов, их истолкование. В других случаях контент-анализ путают с реферированием текстов или с поиском информации в текстовых базах данных.

Появление контент-анализа было реакцией на возникшую потребность в создании объективных методов анализа текстов, результаты которых не зависели бы ни от личности исследователя, ни от того, где и когда эти исследования проводятся. Т.е. требовалось найти такие методы оценки текстов, которые не вызывали бы разногласий между исследователями и были воспроизводимы в любое время и в любом месте.

Никто не возражает против содержательного анализа текстов, их истолкования и пр. Просто не следует называть это контент-анализом, который изначально он задумывался именно как *строгий метод оценки текстов.*

Одним из определений контент-анализа является следующее: *"Контент-анализ - это методика выявления частоты появления в тексте определенных интересующих исследователя характеристик, которая позволяет ему делать некоторые выводы относительно намерений создателя этого текста или возможных реакций адресата"* [10].

Когда в качестве наиболее объективной оценки текстов избрали частоту появления в нем различных характеристик, казалось, что оптимальное решение найдено. Вскоре поняли, что не все так просто. Если попросить двух экспертов подсчитать, сколько раз, например, было упомянуто имя президента в конкретном номере конкретной газеты, то скорее всего их ответы совпадут. Причиной расхождений может стать лишь *невнимательность* при подсчете. Но вот если попросить этих же экспертов подсчитать в той же газете количество слов с негативной окраской, то результаты будут явно отличаться. Более того, один и тот же эксперт на одном и том же материале в разные моменты времени даст разные ответы. Причина кроется в *неоднозначности* критериев. Эта проблема стоит настолько остро, что она даже отдельно изучается. Существуют специальные методы оценки надежности результатов ручного контент-анализа, когда можно доверять экспертам, а когда нельзя.

Отдельный вопрос - *трудоемкость* контент-анализа. Имеется интересная методика, позволяющая по тексту объемом от 80 до 150 слов получить достаточно полный психологический портрет автора. Анализируются в основном грамматические характеристики. На ручной анализ одного текста по той же методике уходит от 4 до 6 часов времени. Гораздо хуже обстоят дела, когда приходится оценивать большие массивы текстов, поступающих непрерывно. Ручной контент-анализ становится просто невозможным. Выходом в данной ситуации является разработка компьютерных методов контент-анализа. *Невнимательность* исключена; *неоднозначность* исключена, если критерии приняты; *трудоемкость* решается за счет быстродействия.

К математическим оценкам текстов в компьютерном контент-анализе можно предъявить ряд требований. Во-первых, эти оценки должны сами по себе иметь хорошее математическое

обоснование. Во-вторых, они должны быть просты, понятны и легко интерпретируемы даже людьми далекими от математики. Лишь в этом случае методы контент-анализа получат широкое распространение и применение в гуманитарных исследованиях. В-третьих, они должны допускать удобное наглядное представление не только в виде таблиц чисел, но и в виде графиков и диаграмм. Последнее просто в иной форме выражает требование к удобному интерфейсу компьютерных программ, позволяющему отображать данные как в дискретной, так и в аналоговой форме.

Характеристиками или элементами содержания, по отношению к которым применяется процедура подсчета, могут быть отдельные слова, словосочетания, предложения, абзацы, тексты. При этом сами характеристики никогда не являются самоцелью. Они интересны лишь в той степени, в какой являются индикаторами происходящего во внеязыковой реальности. В этом заключается существенное отличие контент-анализа от методов квантитативной лингвистики, от методов статистического изучения языка.

Оценки частот

В контент-анализе самыми бедными по содержанию и в то же время самыми фундаментальными являются простые оценки частот. Примем следующее обозначение

$f(c,t)$ - частота встречаемости характеристики c в тексте t.

В качестве примера рассмотрим частоту (количество) упоминания фамилии конкретного политика в конкретном СМИ (газете). Если речь идет о частоте упоминания в отдельном номере газеты, то практически никаких выводов сделать из этого нельзя. Совсем другое дело, если отслеживать частоты на протяжении определенного отрезка времени и сопоставлять их с поступками этого политика. Отсюда можно прийти к выводу о том, что в поведении данного политика привлекает внимание журналистов анализируемого издания. Можно подсчитывать частоту упоминания политика не в отдельных номерах газеты, а помесячно, и сопоставлять ее не с поступками, а с регулярно публикуемыми рейтингами политических деятелей. Это явится подходящим материалом для исследования на тему, как влияет и

влияет ли частота упоминания политика в СМИ на его рейтинг. Гораздо больше информации даст одновременный подсчет частот упоминания не одного, а нескольких политиков. Появляется возможность сравнивать их между собой. В этом случае, например, корреляции частот может послужить основанием для более глубокого изучения общего в поведении анализируемых политиков.

Отдельные слова, как элементы содержания, являются частным случаем того, что в контент-анализе называется категорией. *Категория - это множество слов, объединенных вместе по тому или иному признаку.* Так, например, в качестве категории *ЖИЛЬЕ* может выступать группа синонимов *{берлога, дом, жилище, жилье, логово, логовище, обиталище, обитель}*. Другими примерами могут быть категории агрессивно окрашенной лексики *АГРЕССИВНОСТЬ={бить, бушевать, грозить, назло, одолеть, погром, рычать,...}* и позитивно окрашенной лексики *ПОЗИТИВ= {благодарность, бодрый, вкусный, добро, нежный, няня, теплый, шутка, юмор, ясный,...}*. Частота упоминания в тексте некоторой категории подсчитывается как сумма частот входящих в нее слов, т.е. если K - категория, то

$$f(K,t)=\sum_{w \in K} f(w,t).$$

Логической операцией, лежащей в основе создания категории, является определение через абстракцию. Вовсе не обязательно категория должна задаваться посредством заранее фиксированного списка слов. Иногда гораздо удобнее задать ее операционально. Примером такой категории может быть категория глаголов прошедшего времени. Определение принадлежности к ней будет заключаться не в сопоставлении с фиксированным списком слов, а в распознавании грамматических признаков глагола прошедшего времени.

Более сложными являются категории, состоящие не просто из отдельных слов, а из целых словосочетаний. Например, категория *МОРЕ={Черное море, Средиземное море, Красное море, Балтийское море,...}*. Контент-анализ с использованием категорий позволяет оценивать тексты на более высоком абстрактном уровне. Результаты, получаемые с их помощью, качественно богаче. Возьмем, например, категории *ПОЗИТИВ*,

НЕГАТИВ, АГРЕССИВНОСТЬ, АРМИЯ, ПОЛИТИКА, ЭКОНОМИКА, РАЗВЛЕЧЕНИЯ, ЗАКОН и подсчитаем частоты их встречаемости в интересующем нас издании на протяжении нескольких месяцев. Затем сопоставим, подсчитаем корреляцию, с ежемесячными рейтингами этого же издания среди различных социально-демографических групп. Положительные и отрицательные коэффициенты корреляции между частотами отдельных категорий и рейтингами подскажут, статьи какой тематики привлекают или отталкивают читателей той целевой группы, на которую рассчитано издание.

Как было сказано ранее, не только слова или словосочетания являются теми элементами содержания, частота которых может интересовать исследователя. Вместо того, чтобы подсчитывать частоту упоминания фамилии политика, можно подсчитывать частоту предложений, в которых упоминается политик. Очевидно, что в общем случае вторая величина будет меньше первой. Можно подсчитывать частоту абзацев, обладающих определенными признаками. Более крупными элементами являются целые тексты - статьи и книги. Например, подсчет частоты статей различной тематики позволяет делать выводы о редакционной политике издания. Аналогичный подсчет тематики книг, поступающих в научную библиотеку, позволяет судить о тенденциях в развитии науки, перспективных направлениях исследований и т.д.

Условные частоты

Простые частоты являются не самой подходящей оценкой текстов. Проблемы с ними могут возникнуть в том случае, если мы захотим сравнить разные по длине тексты. Например, пусть в некотором тексте t_1 длиной в 1000 слов категория *НЕГАТИВ* встречается с частотой 20, а в тексте t_2 длиной в 10000 слов - с частотой 100. Является ли пятикратная разница частот достаточным основанием для утверждения, что текст t_2 окрашен более негативно, чем текст t_1? Очевидно, что нет. Для вынесения такого утверждения необходимо сравнивать не простые частоты, а условные, т.е. доли которые составляет категория *НЕГАТИВ* в первом и втором тексте.

Условную частоту характеристики c в тексте t обозначим посредством pr(c,t). Вычисляется она по формуле

$$pr(c,t)=f(c,t)/L(t),$$ где L(t) - длина текста t

В зависимости от того, что принято за элементы содержания, в качестве длины текста может быть взято общее количество в нем слов, количество предложений, количество абзацев и т.д. Обычно, если характеристика - это отдельное слово или категория слов, то и в качестве длины текста берется количество слов в нем.

В нашем примере pr(*НЕГАТИВ*,t_1)=20/1000=0,02 больше, чем pr(*НЕГАТИВ*,t_2)=100/10000=0,01. Т.е. более негативно окрашенным является не второй, а первый текст.

Иногда вместо условных частот удобнее использовать оценку процентного содержания. Для этого просто умножают условную частоту на 100 и тем самым получают процентное содержание. Переход от использования простых частот к условным значительно расширяет сферу применимости методов контент-анализа. Если раньше все наши примеры имели дело с текстами одинаковой длины, то теперь это ограничение снято. Теперь мы можем сравнивать разные по длине статьи, разные по объему издания и пр.

Нормы

До сих пор для того, чтобы делать какие-то выводы, нам требовалось оценить как минимум два текста. Затем эти оценки либо сопоставлялись между собой, либо соотносились с некоторыми событиями в реальном мире, и на основании этого делались определенные выводы.

Представим, что перед нами поставлена задача классификации текстов по медицинской и немедицинской тематике. Причем требуется, чтобы это делал не человек, а компьютер. Решение довольно очевидно. Текст должен быть отнесен к медицинским в том случае, если частота встречаемости медицинских терминов в нем существенно выше, чем в обычной речи. Для этого следует сформировать категорию медицинских терминов Km и сопоставить ей условную частоту встречаемости в обычной речи pr(Km,речь), которую назовем *нормой для категории* Km. При анализе конкретного текста t

подсчитывается условная частота pr(Km,t). Если она существенно больше нормы pr(Km,речь), то текст t относят к к медицинской тематике. Аналогичная процедура может быть применена для дальнейшей классификации текстов по различным разделам медицины. Достаточно лишь сформировать соответствующие категории и сопоставить им нормы, но уже не на основании обычной речи, а на основании анализа представительной выборки различных медицинских текстов. Задача по формированию норм облегчается тем, что в настоящее время существует довольно много частотных словарей, относящихся к различным сферам человеческой деятельности, и нормы можно извлекать из них. Нормы можно вычислять и для отдельных людей. Они могут оказать весьма полезны, например, для определения душевного состояния человека. Так превышение в речи относительно личной нормы частоты категории *НЕГАТИВ* может свидетельствовать о том, что человек находится в дурном настроении.

Важно подчеркнуть, что понятие нормы всегда относительно. Для сугубо гражданского человека норма частоты употребления агрессивно окрашенной лексики одна, для профессионального военного - другая. Нормы могут меняться не только от одной профессионально определенной группы людей к другой, но и со временем. Причиной тому служат исторические изменения в жизни общества, отмирание старых идей и появление новых, заимствования из других языков, влияние на лексический состав языка таких факторов как общественная мораль и пр.

Более строго понятие нормы можно определить следующим образом. Имеется некоторое множество текстов T, которые объединены вместе по определенному признаку. Нас интересует норма характеристики c для T. Так как множество текстов T может быть слишком велико или недоступно целиком, то из него берется представительная конечная выборка $V \subseteq T$ и уже для нее вычисляется условная частота pr(c,V). Это и будет принято в качестве нормы характеристики c для T, которую мы обозначим посредством nr(c,T). Норма характеристики c для множества текстов T - это ожидаемая условная частота ее встречаемости в произвольном тексте, принадлежащем данному множеству. Для представления того, как сильно отличается от ожидаемой частота встречаемости характеристики c в конкретном тексте $t \in T$, используются следующие оценки:

pn(c,t,T)=pr(c,t)/nr(c,T) - во сколько раз отличается pr(c,t) от nr(c,T)

pd(c,t,T)=[[pr(c,t)-nr(c,T)]/nr(c,T)]*100 - на сколько процентов отличается pr(c,t) от nr(c,T).

Аналитика в первую очередь интересуют те тексты, для которых оценка pn(c,t,T) существенно отличается от 1, или же оценка pd(c,t,T) существенно отличается от 0. При этом дополнительного уточнения тербует термин *существенно отличаться*. На помощь приходит аппарат математической статистики. Обычно считают, что характеристика c имеет в тексте t биномиальное распределение с вероятностью nr(c,T). Пусть реально в тексте t характеристика c встретилась pr(c,t)*L(t) раз в то время как ожидалось nr(c,T)*L(t). Исходя из свойств биномиального распределения легко подсчитать, насколько мала вероятность того, что для произвольного текста t_i абсолютная величина abs(pr(c,t_i)-nr(c,T))*L(t_i) ≥ abs(pr(c,t)-nr(c,T))*L(t). Обычно, если вычисленная таким образом вероятность не превышает порога 0,05 (или 0,01), считается, что отклонение реальной частоты от ожидаемой существенно, т.е. не является случайным.

На практике гораздо чаще используют оценку, вычисляемую по формуле:

z(c,t,T)=[pr(c,t)-nr(c,T)]/SQRT[pr(c,t)*(1-pr(c,t))/L(t)]

Это разница двух условных частот, нормированная по стандартному отклонению. Ее имеет смысл использовать лишь в том случае, если pr(c,t)*(1-pr(c,t))*L(t)≥25. Эта оценка хорошо известна психологам и социологам. Именно с ее помощью обосновываются методы вычисления баллов многих психологических тестов. Если z(c,t,T)≥1,96, то мы сразу можем сказать, что вероятность данного события не превышает 0,05. Если же z(c,t,T)≥2,58, то вероятность этого события еще меньше и не превышает 0,01. Из формулы видно, что данная оценка прямо пропорциональна корню квадратному из длины текста t. Именно поэтому ее можно использовать для определения того, что данное событие не является случайным, но не для оценки того, насколько велико отклонение реальной частоты от

ожидаемой. К сожалению, многие психологи и социологи не различают этого и потому их выводы очень далеки от научности. В применении к методам психологического тестирования замечательную критику по этому вопросу дал А.Г.Шмелев.

Контекстный анализ

Основная идея контекстного анализа заключается в том, что анализу подвергается не весь текст, а лишь некоторая выборка из него, являющаяся контекстом употребления характеристики c. Есть много способов задать контекст. Например, для слова (характеристики) w в качестве его контекста мы можем взять все предложения (абзацы, статьи, книги), в которых оно встречается. Вместо предложений мы можем считать контекстом по одному или более слов слева и справа от каждого вхождения w в текст.

Если текст t рассматривать как множество предложений, а предложение s рассматривать как множество слов, то контекст категории C в тексте t можно определить как

$$ctx(C,t)=\{s-\{w\} \mid w\in C,\ w\in s,\ s\in t\}.$$

Выделенный контекст может анализироваться как самостоятельно, так и относительно основного текста. Во втором случае основной текст служит источником норм, которые затем используются при анализе контекста. Т.е. во втором случае для произвольной категории K мы интересуемся условной частотой pr(K,ctx(C,t)) и сравниваем ее с нормой nr(K,t), вычисляемой как pr(K,t-{C}), где $t-\{C\}=\{s-\{w\} \mid w\in C,\ s\in t\}$

Дополнительно к этому мы можем выделить множество слов

$$col(C,t)=\{w \mid pr(w,ctx(C,t)) \text{ существенно больше } pr(w,t-\{C\})\}$$

В англоязычной литературе по контент-анализу такое множество называется collocation категории C. Отношение *существенно больше* валидизируется с помощью аппарата математической статистики по аналогии с тем, как это описывалось выше. Множество col(C,t) содержит много полезной информации о категории C. Например, col({змея},речь) будет содержать такие слова как *яд, кусать, ползать,*

пресмыкающееся,..., а в col({*Путин*},*СМИ*) войдут слова *Владимир, президент, Кремль, Россия,...*.

Связи категорий

Мы можем интересоваться не только оценками данного текста по отдельных категориям, но и их взаимосвязями.

Любому тексту t, рассматриваемому как последовательность предложений $<s_1,...s_n>$, и категории C может быть сопоставлен булев вектор $b(t,C)=<v_1,...v_n>$, где $v_i=1$, если для некоторого $w \in C$ имеет место $w \in s_i$, и $v_i=0$ в противном случае. На множестве векторов легко определить логические операции. Для двух векторов $b(t,C_i)=<v_1,...v_n>$ и $b(t,C_j)=<u_1,...u_n>$ они определяются следующим образом

$$b(t,C_i)\&b(t,C_j)=<\min(v_1,u_1),...,\min(v_n,u_n)> \text{- конъюнкция}$$

$$b(t,C_i)\vee b(t,C_j)=<\max(v_1,u_1),...,\max(v_n,u_n)> \text{- дизъюнкция}$$

$$\neg b(t,C_i)=<1-v_1,...,1-v_n> \text{- отрицание}$$

Затем на множестве векторов можно ввести логические отношения *совместности, противоречия, подчинения* и пр. Очевидно, что таким образом задается некоторая логическая модель предметной области, о которой идет речь в тексте, или же модель когнитивной карты, присущей автору текста. Дальнейшее изучение этих моделей проводится с использованием аппарата классической, многозначной или вероятностной логики высказываний.

Очевидно, что мы можем применить к анализу взаимосвязи категорий внутри текста тот же аппарат вероятностной логики, который применили в предыдущей главе к анализу запросов.

Особый интерес представляет анализ и визуализация отношений между категориями с использованием аппарата многомерного шкалирования, кластерного и факторного анализа.

Определим на множестве категорий (булевых векторов, сопоставленных категориям) функцию близости. Для каждого вектора $b(t,C_i)=<v_1,...v_n>$ вычисляется оценка

$$p_i = \sum v_j/n \quad j=1,\dots,n$$

Тогда коэффициент корреляции для булевых векторов вычисляется следующим образом

$$cor(C_i,C_j) = (p_{i\&j} - p_i*p_j) / sqrt(p_i*(1-p_i)*p_j*(1-p_j)),$$

а функцию близости можно определить как

$$d(C_i,C_j) = 1-cor(C_i,C_j)$$

Также в качестве оценки близости двух категорий часто используется метрика Хемминга, определяемая посредством формулы

$$h(C_i,C_j) = p_i + p_j - 2*p_{i\&j}$$

Контент-мониторинг

Если анализу подвергается массив упорядоченных во времени текстов, поступивших из одного источника, речь идет уже не о простом контент-анализе, а о контент-мониторинге текстовой информации. В этом случае, появляется дополнительная возможность применить математический аппарат многомерного регрессионного анализа и аппарат анализа временных рядов.

Так, например, контент-мониторинг размещенных в сети Интернет пресс-релизов РАО ЕЭС позволил обнаружить закономерности, связывающие различные психолингвистические характеристики текстов с последующими биржевыми изменениями курса акций компании. Применение этих же закономерностей к анализу пресс-релизов компании ENRON, размещенных на ее Интернет-сайте, позволило обнаружить неблагополучие в делах компании задолго до наступившего осенью 2001 года банкротства. То, чего не заметили аудиторы, было обнаружено с использованием методов контент-мониторинга.

1. КОМБИНИРОВАННАЯ ЛОГИКА ЗАПРОСОВ

Логика запросов к сети Интернет представляет собой дуал позитивного фрагмента классической логики. В ней нет общезначимых формул, но есть противоречивые. По этой причине ее удобно представить в виде комбинированного исчисления высказываний и событий [8].

Def.1 Язык

1. Множество событийных переменных Var = {p, q, r, ...};
2. Двухместные функциональные символы $\cap$, $\cup$,#;
3. Формулообразующий оператор θ;
4. Логические связки &, $\vee$, $\supset$, $\neg$.

Def.2 Термы

1. Если $p \in Var$, то $p \in Term$;
2. Если $a \in Term$, $b \in Term$, то $a \cap b \in Term$, $a \cup b \in Term$, $a\#b \in Term$;
3. Ничто другое термом не является.

Def.3 Формулы

1. Если $a \in Term$, то $\theta a \in Frm$;
2. Если $A \in Frm$, $B \in Frm$, то $A \& B \in Frm$, $A \vee B \in Frm$, $A \supset B \in Frm$, $\neg A \in Frm$;
3. Ничто другое формулой не является.

Def.4 Модель

Моделью будем называть пару M=<W, П>, где W – множество возможных миров, а П – семейство его подмножеств, замкнутое относительно пересечения, объединения и относительного дополнения.

Пусть $Val = \Pi^{Var}$. Для фиксированной модели M определим v(a) - значение терма a в модели M при приписывании значений переменным v.

Def.5

1. $v(a \cap b) = v(a) \cap v(b)$;
2. $v(a \cup b) = v(a) \cup v(b)$;
3. $v(a\#b) = v(a) \setminus v(b)$.

Для фиксированной модели M Определим отношение $(v,x)\models A$ – «формула A истинна в модели M в мире x для приписывания значений переменным v».

Def.6

1.	$(v,x)\models \theta a$	$\Leftrightarrow$	$x\in v(a)$
2.	$(v,x)\models A\&B$	$\Leftrightarrow$	$(v,x)\models A$ и $(v,x)\models B$
3.	$(v,x)\models A\vee B$	$\Leftrightarrow$	$(v,x)\models A$ или $(v,x)\models B$
4.	$(v,x)\models A\supset B$	$\Leftrightarrow$	$(v,x)\not\models A$ или $(v,x)\models B$
5.	$(v,x)\models\neg A$	$\Leftrightarrow$	$(v,x)\not\models A$

Определим отношение $M,x\models A$ – «формула A истинна в мире x в модели M».

Def.7

$M,x\models A \Leftrightarrow$ для всякого $v\in Val$ имеет место $(v,x)\models A$.

Определим отношение $M\models A$ – «формула A истинна в модели M».

Def.8

$M\models A \Leftrightarrow$ для всякого $x\in W$ имеет место $M,x\models A$.

Определим отношение $\models A$ – «формула A общезначима».

Def.9

$\models A \Leftrightarrow$ для всякой модели M имеет место $M\models A$.

Аксиомы

1. Аксиомы классической логики высказываний;
2. $\theta(a\cap b)\equiv\theta a\&\theta b$;
3. $\theta(a\cup b)\equiv\theta b\vee\theta a$;
4. $\theta(a\#b)\equiv\theta a\&\neg\theta b$;

Правило вывода:

R.1 $\vdash A,\ \vdash A\supset B \Rightarrow\ \vdash B$.

Имеют место следующие теоремы.

Теорема о непротиворечивости. $\vdash A \Rightarrow \models A$.

Теорема о полноте. $\models A \Rightarrow \vdash A$.

Построенная логика может быть представлена в виде аналитических таблиц [7]. Для этого достаточно добавить к стандартной формулировке классической логики следующие шесть правил редукции:

1. $\{T\theta(a \cap b), \Sigma\} \Rightarrow \{T\theta a, T\theta b, \Sigma\}$
2. $\{F\theta(a \cap b), \Sigma\} \Rightarrow \{F\theta a, \Sigma\}, \{F\theta b, \Sigma\}$
3. $\{T\theta(a \cup b), \Sigma\} \Rightarrow \{T\theta a, \Sigma\}, \{T\theta b, \Sigma\}$
4. $\{F\theta(a \cup b), \Sigma\} \Rightarrow \{F\theta a, F\theta b, \Sigma\}$
5. $\{T\theta(a \# b), \Sigma\} \Rightarrow \{T\theta a, F\theta b, \Sigma\}$
6. $\{F\theta(a \# b), \Sigma\} \Rightarrow \{F\theta a, \Sigma\}, \{T\theta b, \Sigma\}$

Условия замыкания таблицы остаются теми же, что и для классической логики.

2. АЛГОРИТМ ПОСТРОЕНИЯ АНАЛИТИЧЕСКИХ ЗАПРОСОВ

В этом приложении будет подробно описан алгоритм формирования аналитических запросов к сети Интернет. Аналитическими они названы по той причине, что, во-первых, ответы на них содержат не фактическую, а аналитическую информацию, и, во-вторых, сам ответ получается не в явной форме, а требует осуществления некоторой аналитической процедуры. Такие запросы к сети Интернет лучше всего производить с помощью поисковых систем AltaVista для английского языка и Яndex для русского. Использование системы Rambler нежелательно по причине неудовлетворительного алгоритма поиска и индексации, который в ней реализован.

Аналитические запросы производятся с целью получить оценку характера и силы ассоциативной связи в сети Интернет между ответами на простые запросы. Например, нас может интересовать, имеется ли ассоциативная связь между ответами на запросы, представленные словами *war* и *petroleum*.

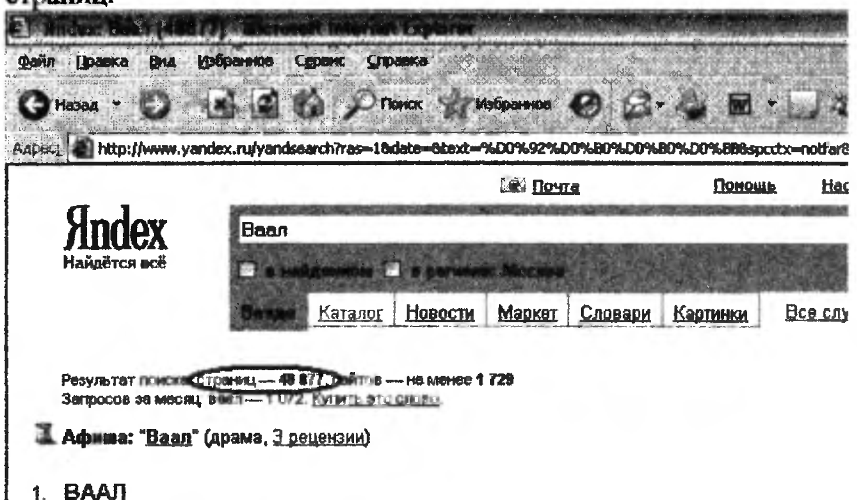
Первый шаг алгоритма заключается в составлении на языке конкретной поисковой системы двух запросов **p** и **q**, отношение между которыми нас интересует. Эти запросы не обязательно должны быть представлены отдельными словами, а могут иметь произвольную степень сложности.

Второй шаг состоит в фиксации контекста поиска. Т.е. запросы производятся не относительно всей сети Интернет, а относительно тех страниц, на которых, например, упоминается словосочетание *United States*. Для этого на языке поисковой системы составляется третий запрос **u**. Он также может иметь произвольную степень сложности, но предпочтение следует отдавать более простым запросам.

Третий шаг. С помощью программы Excel создается таблица следующего вида:

	A	B	C	D	E	F	G	H	I
1	u	u&p	u&q	u&p&q	u&-p&q	u&p&-q	u&-p&-q	X2	P(u&q&p)/(P(u&q)*P(u&p))
2									
3									

Четвертый шаг. В поисковой системе Яndex выбирается опция *Расширенный поиск* (*Advanced Search* в AltaVista) и последовательно производятся четыре запроса вида: **u, u&p, u&q, u&p&q**. Символ & соответствует естественноязыковому союзу 'и', в разных поисковых системах он может быть представлен по-разному. Очевидно, что от исследователя требуется хорошее знакомство с языком запросов поисковой системы. После получения ответа на каждый из запросов в таблицу Excel заносится информации о количестве найденных страниц.



В результате таблица примет следующий вид:

	A	B	C	D	E	F	G	H	I
1	u	u&p	u&q	u&p&q	u&-p&q	U&p&-q	u&-p&-q	X2	P(u&q&p)/(P(u&q)*P(u&p))
2	w	x	y	z					
3									

Пятый шаг. В ячейки таблицы заносятся следующие формулы:

E2→ C2-D2
F2→ B2-D2
G2→ A2-B2-C2+D2
H2→ (D2*A2-C2*B2)*(D2*A2-C2*B2)/ (B2*C2*(A2-B2)*(A2-C2))
I2→ D2*(A2-B2-C2+D2)/((C2-D2)*(B2-D2))
J2→ D2/C2
K2→ D2/B2

Шестой шаг. Определяем уровень значимости. Значение в ячейке H2 сравнивается с числами, занесенными во вторую строку следующей таблицы:

α=0,1	α=0,05	α=0,01	α=0,001
2,71	3,84	6,63	10,83

Выбирается наибольшее число, не превосходящее числа в ячейке H2. Величина α является соответствующим уровнем значимости. Например, если в ячейке H2 стоит число 8,71, то уровень значимости α будет равен 0,01. На практике следует обращать внимание на уровни значимости $\alpha \leq 0{,}05$.

Седьмой шаг – определение характера и силы связи. Если число, стоящее в ячейке I2 больше/меньше единицы, то между двумя анализируемыми запросами имеется ассоциативная/диссоциативная связь, т.е. страницы,

одновременно удовлетворяющие двум запросам, встречаются чаще/реже, чем если бы это имело место в силу чисто случайных причин.

Если производятся аналитические запросы для последовательных временных интервалов, то для них повторяются шаги 4-7. На шаге 5 формулы просто перетаскиваются или копируются в ячейки, стоящие ниже.

Предложенный алгоритм может быть легко реализован разработчиками поисковых систем в качестве дополнительного сервиса.

3. ТЕХНОЛОГИЯ ПРОГНОЗА

В книге Х.Хекхаузена "Мотивация и деятельность" есть параграф под названием *"Констелляции мотивов власти, достижения и аффиляции".* В нем описывается экспериментально обнаруженная связь между выраженностью различных мотивов у руководства фирмы и различными параметрами их (фирм) экономического развития.

"Одну из изучавшихся под этим углом зрения групп составляли люди, занимающие руководящие посты в промышленности. ...Приведенные данные ... позволяют предположить, что оптимальный для экономического роста организационный климат должен складываться, когда руководящие административные потсы занимают люди с высоким мотивом власти, сочетающимся с высоким мотивом достижения и низким мотивом аффилиации...Ту же констелляцию высоких мотивов достижения и власти с низким мотивом аффилиации Кок [S.E.Kock] установил несколько необычным образом. Основываясь на объясненных постфактум результатах работы ряда крупных предприятий, он предсказал их дальнейшую судьбу и проверил свое предсказание через 10 лет.

...Корреляции между значениями отдельных мотивов (а также констелляция "достижение + власть - аффилиация", Д+В-А) руководства фирмы и 5 показателями экономического развития приведены в таблице...Нетрудно видеть, что показатель констелляции (Д+В-А) коррелирует с экономическими показателями более сильно, чем отдельно взятые мотивы достижения, влапсти или аффилиации (последний с обратным знаком). (Стоит добавить, что с увеличением мотива аффилиации уменьшается также объем кредитов).

...Коэффициенты корреляции 5 параметров экономического развития 15 трикотажных предприятий за 1954-1961 гг. и показателей силы мотивов, а также их мотивационной констелляции у руководителей фирм [S.E.Kock "Foretagsledning och motivation", p.215]"

Параметры	Величина мотивов руководства фирмы			
	Достижение (Д)	Власть (В)	Аффилиация (А)	Д+В-А
Совокупная стоимость продукции	0,39	0,49*	-0,61**	0,67**
Количество рабочих мест	0,41	0,42	-0,62**	0,66**
Объем оборота	0,46*	0,41	-0,53*	0,60*
Совокупный объем капиталовложений	0,63*	-0,06	0,20	0,45*
Прибыль	0,27	0,01	-0,30	0,34

Понятно, что данные результаты имеют большое прогностическое значение. Было бы очень интересно, если бы подобные результаты мы могли извлекать из анализа информации, содержащейся в сети Интернет.

Целью настоящей работы как раз и является демонстрация *объективной* связи между психолингвистическими оценками публикуемых в сети Интернет пресс-релизов и состоянием дел компании, представленном в виде курса ее акций.

Временные ряды и корреляции

Людей всегда интересовал поиск закономерностей в окружающей их природе. Если есть два ряда данных

x1, x2, x3,..
y1, y2, y3,...

то в каких случаях можно говорить о наличии связи между ними? Например, **x** - это рост человека, а **y** - его вес. Очевидно, что по значению одного параметра нельзя предсказать точное значение другого, но в то же время очевидно, что связь между ростом и весом человека имеется. Обычно, чем больше рост, тем больше вес, и чем больше вес, тем больше рост. Для оценки такого рода закономерностей, которые могут связывать значения параметров лишь приблизительно, и был придуман в статистике коэффициент корреляции.

Коэффициент корреляции может принимать значения от -1 до +1. Если он отрицателен, то чем больше (меньше) значение одного параметра, тем меньше (больше) значение другого параметра. Если же положителен, то чем больше (меньше) значение одного параметра, тем больше (меньше) значение другого параметра. Крайние значения -1 или +1 указывают на то, что по значению одного параметра мы можем с абсолютной точностью предсказать значение другого параметра. В случае роста и веса людей коэффициент корреляции положителен, но все-таки меньше единицы, так как связь между параметрами не является однозначной.

Если ряды данных связаны с развитием некоторых процессов во времени, то их называют временными рядами. К ним также применимы методы статистического анализа. Появляются дополнительные тонкости, но они также поддаются учету. Например, нас может интересовать связь между государственным финансированием образования и темпами развития экономики. Если в качестве анализируемых данных мы возьмем соответствующие величины для разных стран, то получим коэффициент корреляции, связывающий эти два параметра. Но аналогичное исследование можно провести и для отдельно взятой страны, взяв в качестве анализируемых данных ежегодные вложения в образование и ежегодные оценки темпов развития экономики. Во втором случае мы будем иметь дело с анализом временных рядов. Очевидно также, что во втором случае может быть поставлена задача выявления причинной зависимости между анализируемыми параметрами. Особенность здесь заключается в том, что увеличение или уменьшение финансирования образовательной сферы сказывается на темпах развития экономики не сразу, а с задержкой на несколько лет. Знание таких закономерностей позволяет заблаговременно прогнозировать наступление негативных событий и принимать меры для их предотвращения.

Подтвердить гипотезу о существовании причинной зависимости между двумя временными рядами данных можно путем вычисления так называемой кросскорреляции этих рядов. Это набор коэффициентов корреляции для различных временных сдвигов двух рядов друг относительно друга. Например, при вычислении коэффициентов корреляции мы сравниваем финансирование образования с темпами экономического

развития через год, через два, через три и т.д. Или наоборот сравниваем финансирование образования с темпами экономического развития, какими они были год назад, два года назад, три и т.д. В зависимости от того, в какую сторону осуществлен временной сдвиг, проверяются два варианта гипотезы о направленности причинной связи.

РАО ЕЭС. Постановка задачи.

Путем сравнительного анализа цен на акции РАО ЕЭС и регулярно публикуемых пресс-релизов компании определить, существуют ли закономерности, связывающие чисто психологические оценки содержания пресс-релизов с колебаниями курса акций.

Материал для анализа

В качестве материала для контент-анализа были использованы более 1000 пресс-релизов компании РАО ЕЭС, размещенных в сети Интернет в период с июля 1999 по март 2002 года и объединенные по месяцам. Информация о дневных ценах на акции компании была взята из архивов РТС.

Методы анализа

Компьютерный контент-анализ текстов с помощью психолингвистической экспертной системы ВААЛ.

Результаты анализа

Результаты анализа были представлены в виде таблицы коэффициентов корреляции, связывающих психолингвистические оценки пресс-релизов с колебаниями цен акций. Коэффициенты корреляции были вычислены для различных временных сдвигов от 0 до 5 месяцев. Т.е. нас интересовало не только то, как влияло сегодняшнее содержание пресс-релизов на сегодняшние же цены, но и то, как сказывались сегодняшние пресс-релизы на ценах через 1, 2, 3, 4 и 5 месяцев.

Сообщаемое в пресс-релизах определенным образом соотносится с проводимой компанией хозяйственной деятельностью. Реальные результаты этой деятельности

проявляются через некоторое время. Поэтому нулевой сдвиг (нулевая точка) соответствует непосредственной реакции акционеров на сообщаемую информацию, соответствует ожиданиям акционеров. Оценки же через один и более месяцев соответствуют реакции акционеров на реальные результаты хозяйственной деятельности компании, а не на одну лишь информацию.

Таблица коэффициентов корреляции представлена ниже.

Категории	Месяцы					
	0	1	2	3	4	5
Власть	0,39	0,36	0,47	0,22	0,08	-0,03
Желание власти	0,30	0,29	0,47	0,27	0,14	0,02
Страх власти	-0,03	-0,15	-0,25	-0,30	-0,15	-0,17
Достижение	0,13	0,26	0,45	0,42	0,52	0,19
Достижение успеха	0,13	0,29	0,48	0,47	0,55	0,19
Избегание неудачи	-0,03	-0,18	-0,14	-0,29	-0,19	0,01
Аффилиция	0,17	-0,03	0,09	-0,06	-0,16	-0,29
Надежда на поддержку	0,02	-0,06	0,00	-0,22	-0,18	-0,33
Страх отвержения	-0,29	-0,37	-0,38	-0,29	-0,29	-0,07
Физиология	-0,06	-0,16	-0,30	-0,19	-0,16	-0,04
Потребность	0,07	-0,06	-0,13	-0,18	-0,19	-0,09
Внешняя потребность	0,10	-0,06	-0,17	-0,20	-0,35	-0,21
Внутренняя потребность	-0,06	-0,02	0,06	0,03	0,37	0,30
Валентность общая	-0,11	-0,17	-0,42	-0,47	-0,47	-0,32
Положительная валентность	0,05	-0,15	-0,44	-0,48	-0,51	-0,36
Отрицательная валентность	-0,18	-0,12	-0,29	-0,32	-0,29	-0,17
Инструментальная деят	-0,03	-0,17	-0,19	0,06	-0,09	-0,15
Обработка	0,06	0,16	0,15	0,09	-0,03	0,14
Трансляция	-0,02	-0,03	0,01	-0,17	-0,11	-0,02
Ретрансляция	-0,03	-0,06	-0,10	-0,18	-0,18	-0,08
Движение	-0,22	-0,27	-0,20	-0,12	-0,15	-0,27
Перемещение	-0,39	-0,36	-0,37	-0,18	-0,27	0,00
Манипуляция	0,25	0,09	0,05	0,25	0,22	-0,01

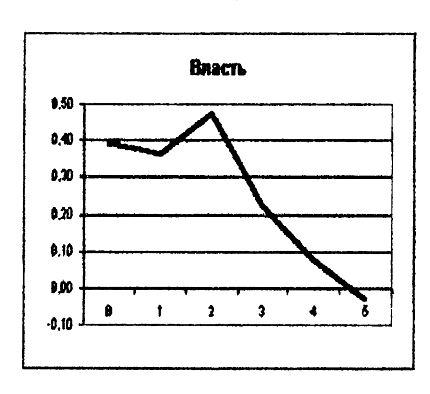
Стремление К	-0,08	-0,24	-0,46	-0,36	-0,18	-0,22
Уход От	-0,40	-0,34	-0,33	-0,04	0,22	0,40
Вверх	0,11	0,23	0,42	0,45	0,13	-0,10
Вниз	0,04	-0,21	-0,37	-0,23	-0,24	-0,25
Отрицание	-0,02	-0,08	-0,24	-0,49	-0,43	-0,37
Женская символика	-0,11	-0,10	0,09	-0,02	0,27	0,21
Мужская символика	0,10	0,08	-0,14	-0,15	-0,32	-0,25
Агрессивность	0,13	0,22	0,21	0,08	-0,02	-0,02
Архетипичность	0,11	0,00	-0,09	-0,19	-0,42	-0,35
Позитив	0,22	0,00	-0,11	-0,28	-0,43	-0,46
Негатив	0,01	-0,05	-0,12	-0,22	-0,24	-0,16
Зрительный канал	0,00	-0,18	-0,27	-0,30	-0,01	0,11
Чувственный канал	-0,04	0,07	-0,13	-0,12	-0,32	-0,36
Слуховой канал	0,02	-0,05	-0,30	-0,45	-0,27	-0,12
Рациональный канал	0,30	0,35	0,55	0,37	0,36	0,01
Несогласие	-0,02	-0,08	-0,25	-0,49	-0,47	-0,42
Согласие	0,29	0,15	0,35	0,11	0,10	-0,06
И	-0,15	-0,11	-0,21	-0,32	-0,09	-0,36
Или	-0,14	-0,09	0,09	0,16	0,48	0,39
Нет	0,05	-0,04	-0,18	-0,41	-0,41	-0,42
Но	-0,06	-0,01	-0,06	-0,14	-0,15	-0,35
Отличие	0,32	0,22	0,08	0,04	-0,15	-0,24
Подобие	0,01	-0,04	-0,08	-0,09	-0,30	-0,05
Д+В-А	0,37	0,46	0,60	0,39	0,32	0,13
Доброжелательность	0,01	0,02	0,21	0,23	0,24	0,15
Интеллект	-0,24	-0,14	0,15	0,29	0,41	0,31
Активность	-0,17	-0,05	0,30	0,39	0,45	0,11
Самоконтроль	-0,27	-0,08	0,06	0,28	0,34	0,37
Независимость	-0,13	0,08	0,19	0,36	0,22	0,14
Раздражительность	0,06	0,09	0,24	0,16	0,00	-0,34
Практичность	0,18	0,24	0,07	0,04	0,07	0,11
Правдивость	-0,06	-0,04	0,03	0,17	0,08	0,11
Доминантность	0,29	0,31	0,23	0,20	-0,06	-0,13
Избалованность	0,25	0,19	0,23	0,29	0,03	-0,19
Деятельность	-0,56	-0,32	-0,18	0,07	0,18	0,26

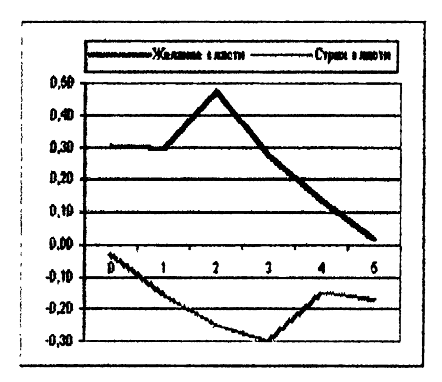
Скрытность	-0,47	-0,44	-0,48	-0,36	-0,17	0,22
Эгоизм	-0,08	0,05	0,03	0,28	0,34	0,23
Утонченность	-0,22	-0,25	-0,08	-0,09	0,14	0,13
Необычность	-0,05	0,00	0,19	0,29	0,21	0,06

Коэффициенты, превышающие по абсолютной величине 0.35, значимы на уровне 5% (с вероятностью 0,95 и выше), а превышающие 0.45, - на уровне 1% (с вероятностью 0,99 и выше).

Мотив Власти

В случае пресс-релизов компании к мотиву *Власти* относится информация о различных управленческих решениях, "связанных с распределением задач, координацией их выполнения, побуждением исполнителей" и пр.





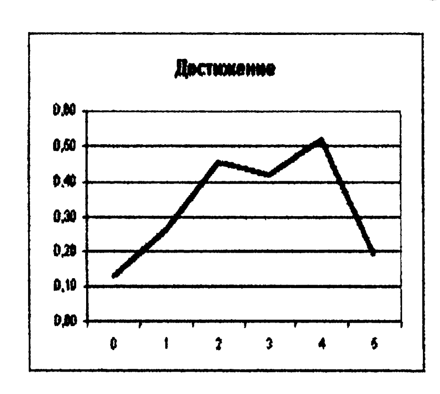
На диаграмме слева мы видим, что ожидания эффектов власти позитивно сказываются на курсе акций. Это следует из того, что коэффициент корреляции в точке 0 равен 0.39. Через месяц коэффициент корреляции остается практически тем же, он равен 0.36. Реальный эффект проявления мотива *Власти* наступает через два месяца, когда коэффициент корреляции достигает значения 0.47. Затем идет его понижение и через 5 месяцев корреляция исчезает.

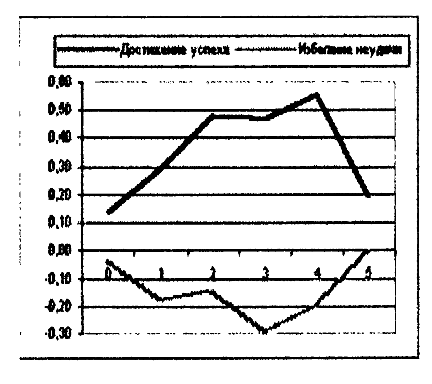
Составляющими мотива власти являются *Желание власти* и *Страх власти*. График *Желания власти* почти в точности

повторяет график самого мотива. Более интересен график *Страха власти*, который в нашем случае можно интерпретировать как нерешительность и бюрократизм в проведении управленческих решений. При нулевом сдвиге *Страх власти* никак не коррелирует с ценами акций. Это значит, что акционеры просто не улавливают никакой нерешительности. Но вот эффекты такой нерешительности все-таки имеются, и своего максимума они достигают через три месяца. *Страх власти* является одной из скрытых переменных, описывающих колебания курса акций. До сих пор эффекты от ее воздействия воспринимались просто как случайные колебания курса.

Мотив Достижения

К мотиву *Достижения* относится деятельность, направленная на достижение результата. Выраженность мотива *Достижения*, как и других мотивов, определяет интенсивность действий, а их эффективность определяют две составляющие этого мотива: *Достижение успеха* (повышает эффективность) и *Избегание неудачи* (понижает эффективность).

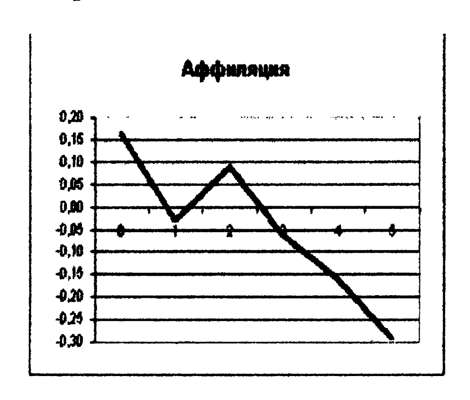


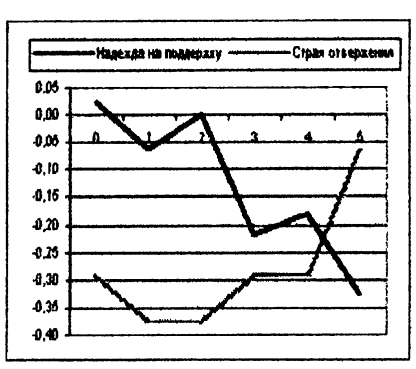


Мотив *Достижения*, как и его составляющие, в точке нулевого сдвига не коррелируют с курсами акций. Это означает, что у акционеров остутствуют соответствующие ожидания. Акционеры вообще не выделяют этого фактора! Реальные эффекты проявления мотива *Достижения* наступают через два, а своего пика достигают через четыре месяца. Для *Достижения успеха* и *Избегания неудачи* ситуация зеркальная.

Мотив Аффиляции

Аффиляция ("потребности в социальной поддержке") - деятельность, направленная на поиск дружеских связей, социальную кооперацию, взаимоподдержку. Она также имеет две составляющие - *Надежду на поддержку* и *Страх отвержения*, соответственно влияющие на успешность деятельности в этом направлении.

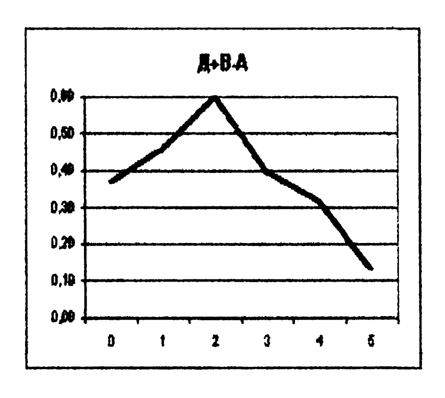


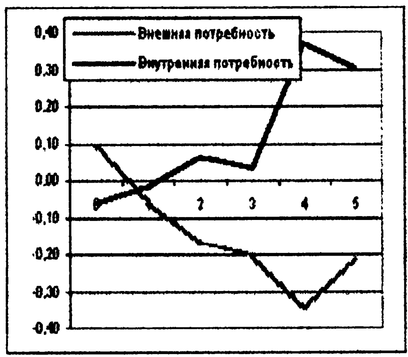


В нулевой точке выраженность мотива *Аффиляции* незначительно (0.17) коррелирует с ценами акций. Акционеры улавливают соответствующую направленность и ожидают, что это положительно (!) повлияет на цены акций. Они не догадываются, что на самом деле влияние будет негативным и его результаты скажутся через 5 месяцев. Цены на акции упадут, но причина этого останется для них [акционеров] сокрыта. Падение так и не будет объяснено.

Д+В-А и Потребность

Согласно Х.Хекхаузену, показатель констелляции мотивов (*Достижение+Власть-Аффиляция*) коррелирует с экономическими показателями более сильно, чем отдельно взятые мотивы *Достижения, Власти* или *Аффиляции* (последний с обратным знаком). Следующая диаграмма демонстрирует это наглядно.

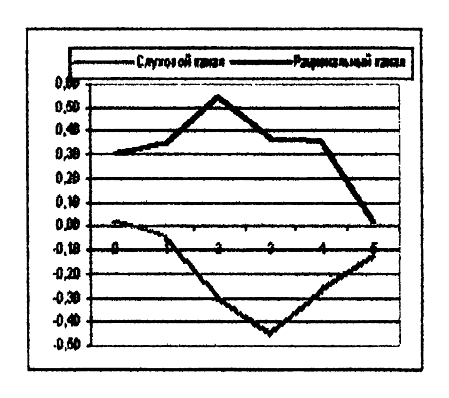


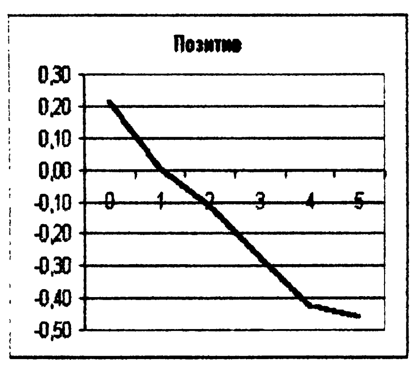


На правой диаграмме показано, как влияет тип мотивации [внутренняя - делаю потому, что хочу; внешняя - делаю потому, что должен] на курсы акций. В нулевой точке ожидания практически отсутствуют, но зато через четыре месяца наступают реальные эффекты. И опять эти эффекты кажутся невесть откуда взявшимися.

Другие оценки

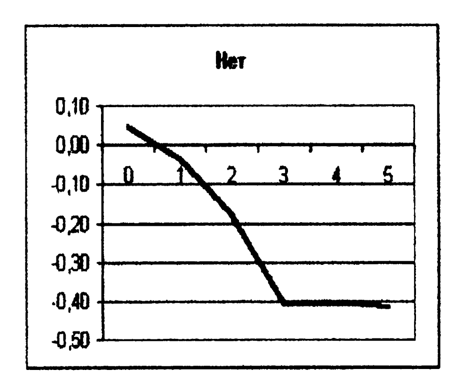
Левая диаграмма показывает, что аппеляция к рассудку, присутствие рациональной аргументации позитивно оценивается акционерами в нулевой точке и эффект от нее [от того, что стоит за этой аргументацией] также положителен. Для слухового канала ситуация зеркальна.

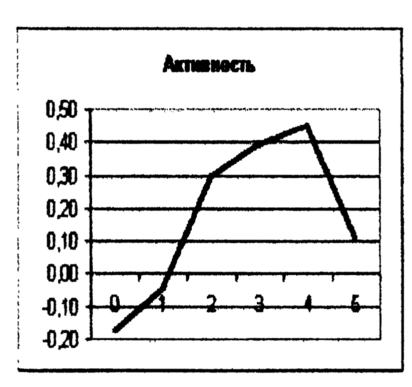




На правой диаграмме показано как связано наличие позитивной лексики с ценами акций. В нулевой момент позитивная лексика положительно влияет на ожидания акционеров, они верят *красивым* словам, но со второго месяца и вплоть до последнего происходит падение коэффициента корреляции. Через пять месяцев он опускается до -0.46. Интересное то, что на эту удочку продолжают ловиться акционеры и у нас и в США.

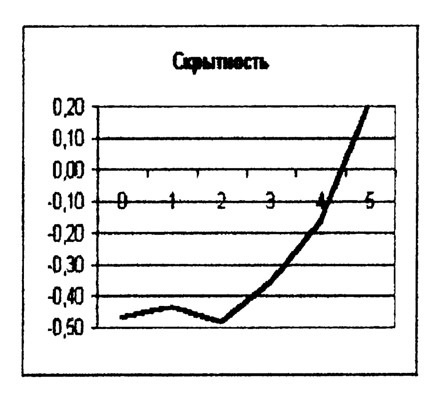
На следующей диаграмме для категории *Нет*, состоящей из слов, представляющих отрицание в русском языке, видно, что увеличение количества таких слов в пресс-релизах практически не замечается акционерами, но всегда связано со значительным падением курса акций через 3 и более месяцев.

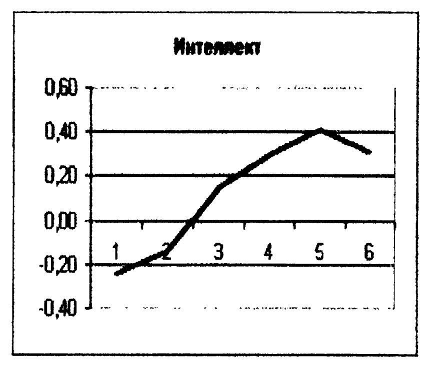




Проявление *Активности* наоборот положительно коррелирует с курсом цен на акции.

Всякое проявление *Скрытности* сразу приводит к падению курса акций, но в долговременной перспективе является оправданным, так как через пять месяцев коэффициент корреляции становится положительным и продолжает свой крутой рост.





Проявление *Интеллекта* с самого начала немножко настораживает акционеров, но в конечном счете приводит к росту курса акций.

Как все это можно использовать?

Мы не предлагаем использовать полученные результаты для того, чтобы предсказать, каким будет курс акций компании РАО ЕЭС через неделю. Для этого существуют другие методы, да это и не интересно.

Полученные результаты позволяют производить глобальную оценку эффективности деятельности компаний. Пример ENRONа как нельзя лучше показал, что даже при внутреннем неблагополучии компании курс ее акций может некоторое время расти и компания даже может быть признана лучшей по итогам года. Но рано или поздно, как говорили диалектики, количество переходит в качество и неожиданно для всех компания объявляет себя банкротом. Теряются огромные состояния, а крайних найти невозможно. Даже аудиторы, на которых пытаются повесить всех собак, оправдываются и справедливо говорят, что оценивают лишь те данные, которые им предоставляют сами компании. Если руководство компании хочет обмануть аудитора, то оно это и сделает. Обманет и аудитора и акционеров.

В этой ситуации косвенный психолингвистический анализ пресс-релизов (не только их) позволяет приподнять завесу таинственности над тем, что действительно происходит с компанией. Как было показано выше, большинство параметров,

по которым диагностируется состояние компании и прогнозируется ее будущее, акционерами не улавливаются. Они не улавливаются для сознательного контроля и самим руководством компании, а потому недоступны для фальсификации. Именно в оценке внутренних тенденций, а не в сиюминутном курсе акций заинтересованы инвестиционные компании и банки для успешного ведения бизнеса. Курс акций может даже расти, но если начала проявляться тенденция падения *Д+В+А*, падения *Внутренней потребности*, роста *Позитива* и *Аффиляции*, то нужно сто раз подумать прежде, чем покупать акции этой компании. А лучше их просто побыстрей продать.

Заключение

На примере РАО ЕЭС мы показали, как реально влияет выраженность различных психологических показателей у высших менеджеров компании на курсы ее акций. Эти закономерности относятся не только к РАО ЕЭС, но имеют универсальных характер, так как их природа заключена не в особенностях данной компании, а в особенностях человеческой психики. Размер компании или ее организационная структура могут повлиять на то, через сколько месяцев проявится эффект от изменения того или иного показателя, но то, что он проявится, неизбежно. Похожие закономерности действуют и в других сферах жизни. Анализируя тексты выступлений высших государственных лиц, можно судить о перспективах будущего развития государства. Анализируя содержание журналов, посвященных различным отраслям промышленности, можно прогнозировать будущее развитие этих отраслей. Много чего можно сделать, и давно уже пора начать это делать.

4. ЛЕТНИЙ БАНКОВСКИЙ КРИЗИС 2004 ГОДА

Целью данного исследования было оценить степень влияния различных средств массовой информации на развитие событий, связанных с летним банковским кризисом 2004 года. Источником информации для анализа послужил Интернет.

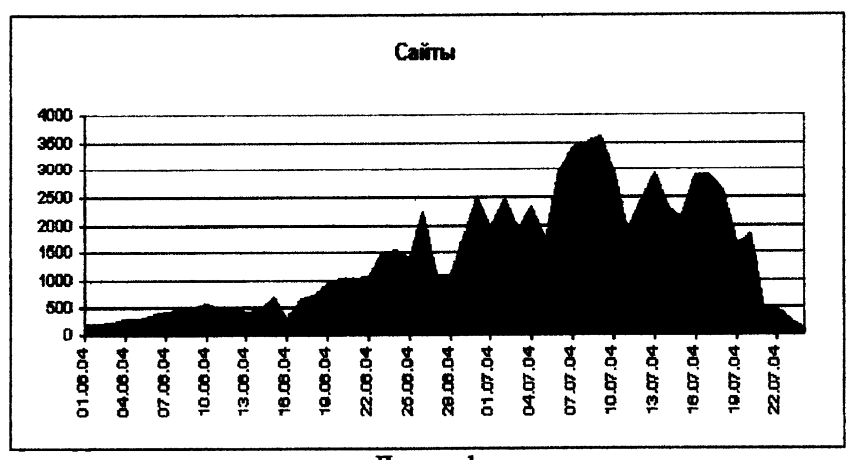
Несостоявшийся банковский кризис вызывает много вопросов. Как могло случиться, что люди поддались мало обоснованной панике и стали в такой спешке снимать свои деньги со счетов в коммерческих банках, что Центробанку пришлось выделить дополнительно более 25 млрд. наличности? Кто его спровоцировал? Кто выиграл? Гарантированы ли мы от повторения подобного?

В настоящей статье будет проведен анализ динамики появления Интернет-публикаций, имеющих отношение к этим событиям. Был выбран период с 1.06.2004 по 24.07.2004. Т.е. 54 дня, на которые и пришелся пик «кризиса» 5-6 июля 2004 года.

Прежде всего, нам необходимо выделить базовые параметры, которые могут служить показателями развития банковского кризиса. В качестве их мы взяли:

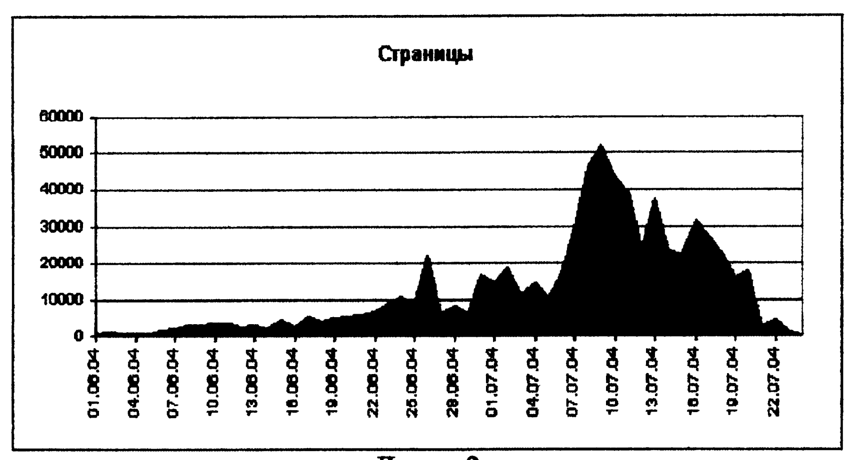
1. количество сайтов по дням на которых появлялись страницы с упоминанием банковского кризиса R[банковский кризис; сайт];
2. количество новых страниц по дням с упоминанием банковского кризиса R[банковский кризис; страница];
3. интенсивность упоминания - R[банковский кризис; страница]/ R[банковский кризис; сайт].

На диаграмме 1 представлена динамика изменения количества сайтов, на которых размещались новые статьи, так или иначе касающиеся «банковского кризиса». Максимум, около 3600 сайтов, пришелся на 9 июля. В последующие дни наметился спад внимания к этому вопросу.



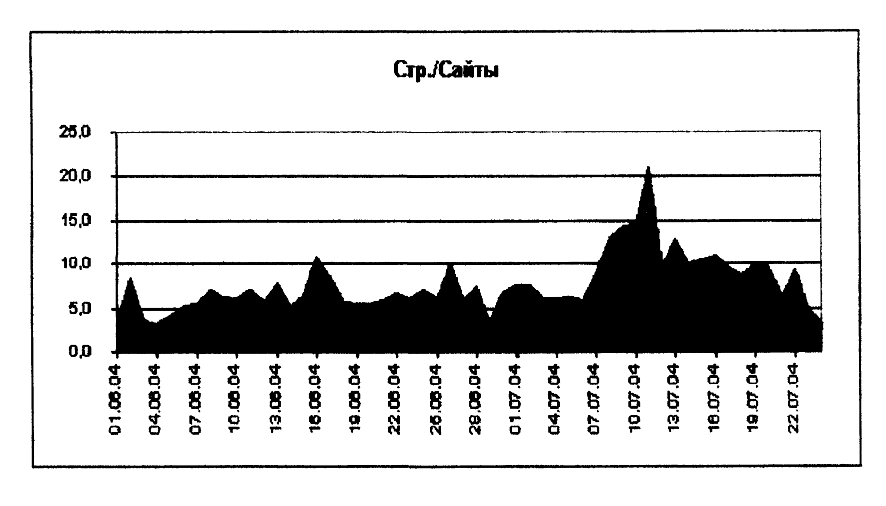
Диагр. 1

На диаграмме 2 показана динамика появления новых документов, размещенных на сайтах в это время. Максимум, около 52 тысяч, приходится также на 9 июля.



Диагр. 2

Следующая диаграмма представляет оценки интенсивности освещения «банковского кризиса» в сети Интернет.



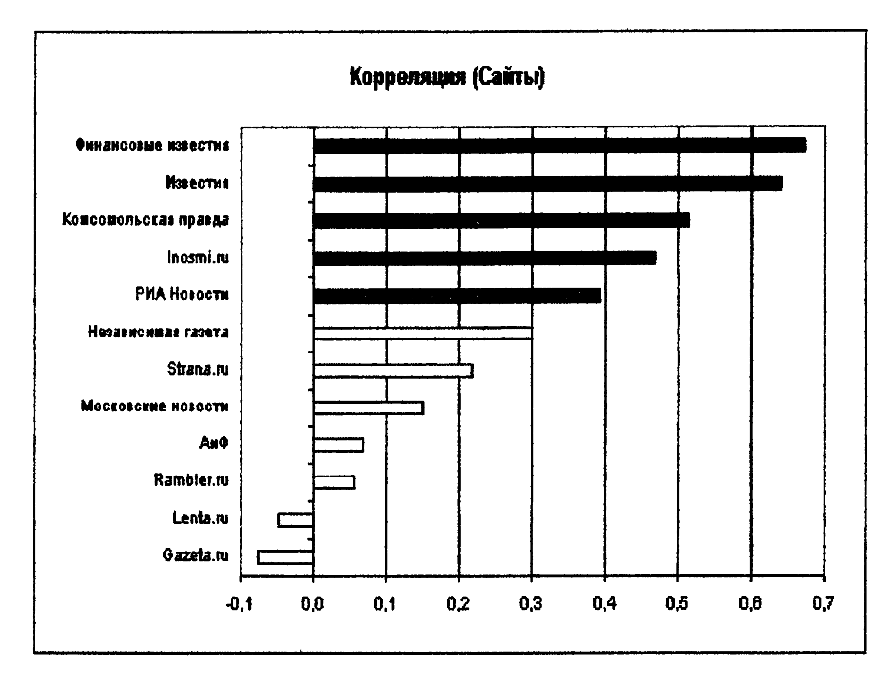
Диагр. 3

Для более пристального анализа нами был выбран ряд СМИ, имеющих, согласно Top 100, наиболее высокий рейтинг посещаемости их Интернет-сайтов.

1. Информационная служба *Strana.ru*
2. Интернет-газета *Gazeta.ru*
3. Интернет-газета *Lenta.ru*
4. Поисковая служба *Rambler.ru* (новостной раздел)
5. *«Известия»* (www.izvestia.ru)
6. *«Финансовые известия»* (www.finiz.ru)
7. *«Комсомольская правда»* (www.kp.ru)
8. *«Независимая газета»* (www.ng.ru)
9. *«Московские новости»* (www.mn.ru)
10. *«Аргументы и факты»* (www.aif.ru)
11. *РИА Новости* (www.rian.ru)
12. Интернет-ресурс *Inosmi.ru*

Для каждого из анализируемых ресурсов был построен график частоты появления новых публикаций по теме «банковского кризиса» и вычислен коэффициент корреляции между этой частотой и оценками диаграммы 1. Т.е. ставилась задача попытаться обнаружить связь между этими параметрами методом сопутствующих изменений. Могли ли публикации в наиболее рейтинговых СМИ спровоцировать *панику вкладчиков,*

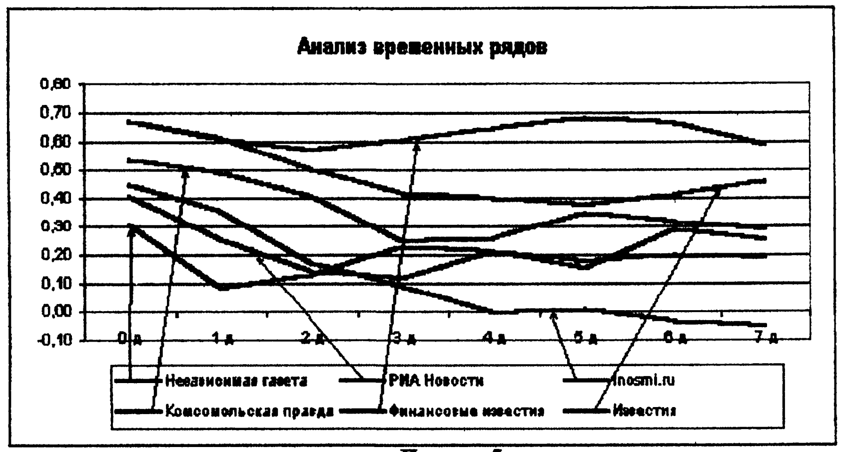
в качестве индикатора которой была взята частота появления сайтов с публикациями по интересующей нас теме? Эти оценки представлены на диаграмме 4. Черным цветом выделены оценки коэффициентов корреляции, значимые на уровне 0,001. Серым - оценки, значимые на уровне 0,05. Белым - статистически незначимые оценки.



Диагр. 4

Априори можно было предположить, что в ситуации отсутствия сильной связи между этими двумя параметрами самыми высокими будут оценки именно у Интернет-изданий *Strana.ru, Lenta.ru, Gazeta.ru и Rambler.ru*, деятельность которых как раз и заключается в оперативном отслеживании происходящих событий и информировании о них в глобальной сети. Т.е. модель описываемых событий представлялась следующим образом. *Среди вкладчиков начинается паника, в сети Интернет на различный сайтах появляется информация о кризисе доверия коммерческим банкам. Параллельно с этим те*

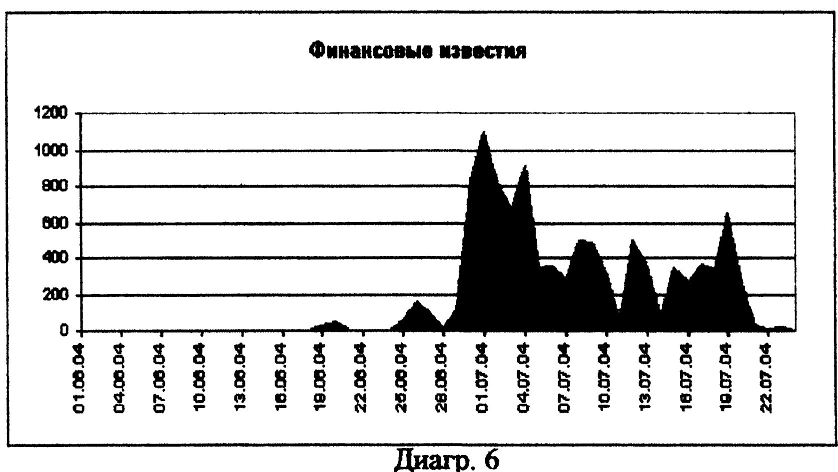
же события отслеживаются и оперативно освещаются упомянутыми выше специализированными электронными СМИ. Корреляция между этими двумя рядами должна быть высокой. Отсутствие ее говорит о наличии других факторов, которые влияли на рост числа проблемных сайтов. Судя по всему, свою роль сыграли публикации на серверах *«Финансовых известий», «Известий», «Комсомольской правды», Inosmi.ru, РИА Новости, «Независимой газеты»*. В этом нас убеждают и результаты анализа временных рядов, представленные на диаграмме 5.



Диагр. 5

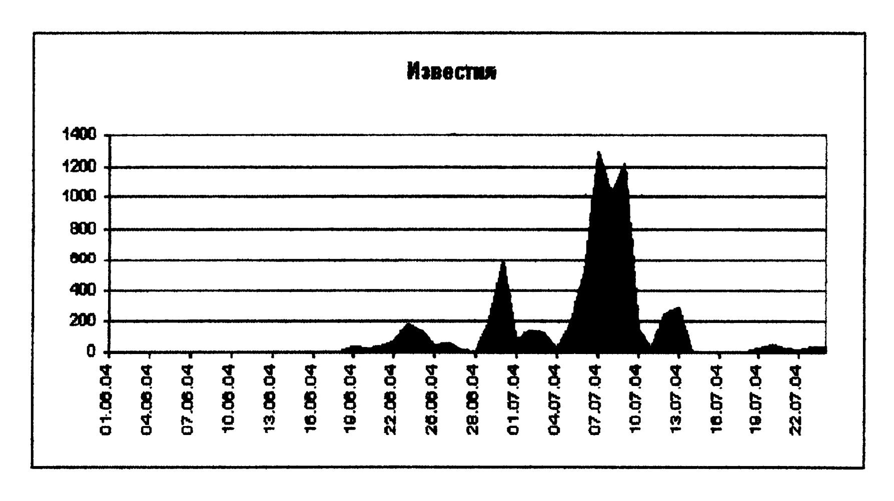
Эти оценки представляют собой коэффициенты корреляции между частотами появления новых публикаций первой шестерки лидеров и частотами появления с временным лагом в 0, 1 ,..., 7 дней аналогичных публикаций на серверах глобальной сети. Из графиков видно, что публикации в *«Финансовых известиях»* и *«Известиях»* продолжали оказывать существенное влияние на появление новых статей по «кризису» в течение 7 последующих дней. Быстрое снижение оценок у *«Комсомольской правды», Inosmi.ru, РИА Новости* и *«Независимой газеты»* говорит о том, что *они скорее откликались на происходящие события, чем активно формировали их.*

Обратимся к графикам частоты появления новых публикаций на анализируемых сайтах.



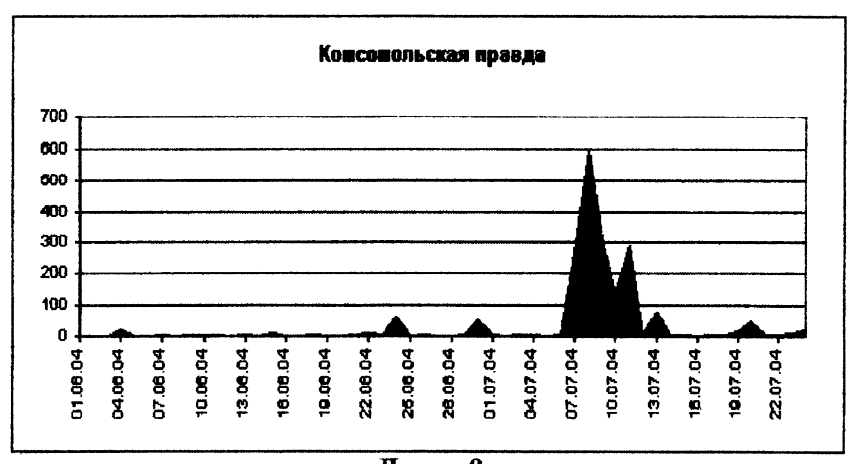
Диагр. 6

Первый пик частоты, 1099 публикаций, приходится на 30 июля, за 6 дней до паники, а второй пик, 913 публикаций, приходится на 4 июля, накануне начала паники. Сомнения вызывают чересчур большие частоты – практически невозможно за один день произвести и разместить столько новых статей. Разгадка заключается в том, что наряду с новыми публикациями, сообщения о «кризисе доверия банкам» появились на главной странице сайта и потому были многократно проиндексированы поисковыми системами. Т.е. ответ на любой запрос к сайту *«Финансовых известий»* подавался в сопровождении информации о «кризисе».



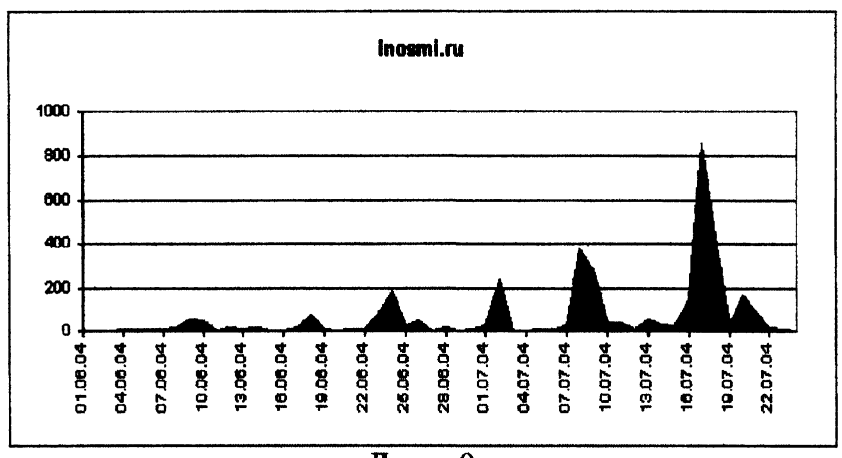
Диагр. 7

Столь же высоки оценки частот и для сайта газеты *«Известия»*. У нее также сообщения о «кризисе» были вынесены на первую страницу. Главное отличие от *«Финансовых известий»* заключается в том, что пик частоты, 1291 публикация, приходится на 7 июля.



Диагр. 8

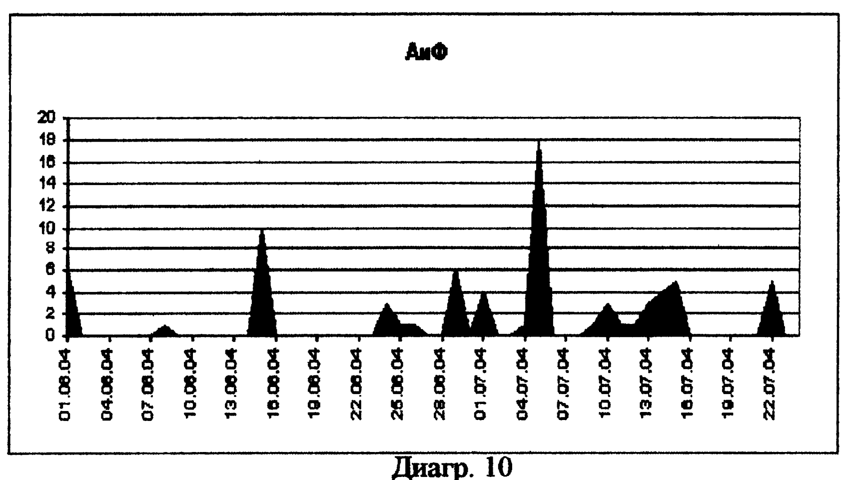
«Комсомольская правда» - 601 публикация 8 июля.



Диагр. 9

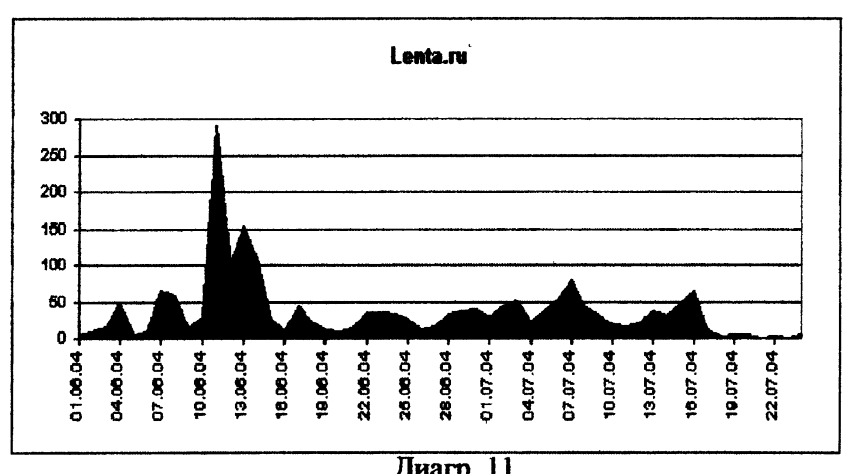
Периодичность появления пиков на сайте *Inosmi.ru* объясняется скорее всего тем, что главное его назначение – периодический обзор зарубежных СМИ.

Трудно сказать, были ли высокие частоты появления новых публикаций на сайтах *«Финансовых известий»*, *«Известий»* и *«Комсомольской правды»* случайными, но аналогичные частоты на сайте *«АиФ»* выглядят вполне естественными – максимум, 18 публикаций, 5 июля.



Диагр. 10

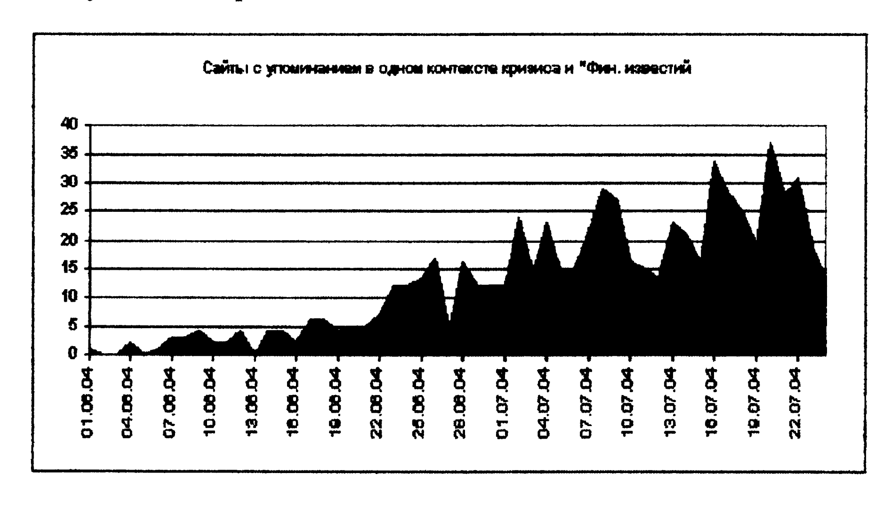
Столь же нормальны частоты новых публикаций на *Lenta.ru.*



Диагр. 11

Основной вывод, который можно сделать из результатов проведенного исследования, заключается в следующем. На поведение вкладчиков во время кризиса могли повлиять публикации в сетевых СМИ. Очевидно, что люди, имеющие сбережения в коммерческих банках, принадлежат к социально-активной части населения. В качестве одного из основных

альтернативных источников информации они используют сеть Интернет. Среди периодических изданий индекс цитируемости сайта газеты *«Известия»*, согласно Яндексу, является вторым по величине (13000) и уступает лишь индексу (15000) *Gazeta.ru*. Среди специализированных изданий индекс цитируемости *«Финансовых известий»* (1900) равен индексу ежедневной аналитической газеты *«RBC Daily»* и значительно превышает индексы других изданий сходной тематики. Во время «кризиса» материалы, публикуемые на сайтах *«Финансовых известий»* и *«Известий»* могли повлиять на решения, которые принимали люди в поисках спасения своих вкладов. Высокие индексы цитирования изданий привели к тому, что подаваемая ими информация стала доминирующей. Мнения, высказанные на страницах *«Финансовых известий»* активно транслировались другими популярными Интернет-ресурсами. В этом убеждает следующая диаграмма.



Диагр. 12

Для того, чтобы получить представление о содержании публикаций *«Известий»* и *«Финансовых известий»*, приведем лишь некоторые цитаты из их статей:

«Где хранить деньги во время банковского кризиса?»
«Россия на пороге беспрецедентного банковского кризиса?»

«Агентство Moody's пересматривает рейтинги 22 российских банков».

«Кризис может изменить динамику цен на жилье».

«Спираль кризиса раскручивается. Власти отрицают очевидное».

«Будет ли банковский кризис в России? Да, случится в 2004 году».

«Как минимум половине из 1300 ныне действующих банков уготована участь вынужденного банкротства или слияния с другими банками».

«Эксперты: гибель Содбизнесбанка и "Кредиттраста" провоцирует банковский кризис». «Отзыв лицензии у Содбизнесбанка и заявление о самоликвидации банка "Кредиттраст" вызвали нестабильность на рынке, которая может перерасти в полномасшабный....»

«Когда случается банковский кризис? Тогда, когда все клиенты банка приходят в банк с одной целью - забрать свои деньги. И не важно, вкладчики это делают или фирмы, держащие в банке свои счета».

«Пострадала надежность еще двух российских банков».

«Кризис из межбанковского перерастает в общеэкономический».

«Чем закончится банковский кризис? Скоро все нормализуется. Исчезнут мелкие и слабые банки».

«Банковский кризис. Низы должны объяснить верхам, как жить дальше».

«Следующий из этого практический вывод - до февраля никакого банковского кризиса не будет. Зато будет серьезное ухудшение финансовых результатов банков во втором полугодии по сравнению с первым. А уже потом у отдельных банков может случиться кризис - если их клиенты от объявленных результатов так расстроятся,..»

«Кризис недоверия. Проблемные банки должны готовиться к худшему».

«Во вторник, 29 июня, еще один банк "средней руки надежности", обладающий довольно широкой сетью...»

«Из кома слухов под названием "банковский кризис" обозначился еще один конкретный - и немалых размеров по российским меркам - фигурант: Гута-банк. 6 июля офисы Гута-банка с раннего утра...»

«Банковский кризис, наличие которого власти продолжают отрицать, продолжает развиваться вне зависимости от их

заверений в обратном. Нельзя сказать, что банкиры совсем уж не виноваты в кризисе. Собственно говоря, вся политика коммерческих банков способствовала тому...»

Выбора у людей не оставалось – необходимо идти и срочно забирать свои деньги.

Остается лишь ответить на вопрос о том, кто нажился на этой панике. Ответ очевиден – ГОСУДАРСТВО. Подтверждением тому служит следующая диаграмма, на которой представлены результаты сравнительного анализа двух временных рядов – интенсивности освещения в сети Интернет «банковского кризиса» и величиной капитализации Сбербанка РФ по данным РТС. Коэффициенты корреляции между двумя рядами вычислялись для временных лагов в 0, 1 ,..., 7 дней.



Диагр. 13

Хорошо видно, что повышение/понижение интенсивности освещения «банковского кризиса» в сети Интернет приводило с небольшой временной задержкой к увеличению/снижению капитализации Сбербанка РФ. Для временного лага в 6 дней величина коэффициента корреляции необыкновенно высока и составила 0,79. Т.е. сила связи приблизилась к функциональной.

ЛИТЕРАТУРА

1. *Гаек П., Гавранек Т.* Автоматическое образование гипотез: математические основы общей теории. М.: Наука, 1984.
2. *Есенин-Вольпин А.С.* Анализ потенциальной осуществимости // Философия. Логика. Поэзия. Защита прав человека: Избранное. М., 1999.
3. *Кайберг Г.* Вероятностная и индуктивная логика. М.: Прогресс, 1978.
4. *Карпенко А.С.* Многозначные логики. М.: Наука, 1997.
5. *Карпенко А.С.* Введение в многозначную пропозициональную логику: Учеб. пособие. М., 2003.
6. *Костюк В.Н.* Подтверждение и принятие гипотезы // Индуктивная логика и формирование научного знания. М., 1987.
7. *Костюк В.Н.* Элементы модальной логики. Киев: Наукова думка, 1978.
8. *Смирнов В.А.* Утверждение и предикация. Комбинированные исчисления высказываний и событий // Логика и системные методы анализа научного знания. М., 1986.
9. *Тюрин Ю.Н., Макаров А.А.* Анализ данных на компьютере /Под. ред. В.Э.Фигурнова. М.: ИНФРА-М, 2003.
10. *Федотова Л.Н.* Анализ содержания – социологический метод изучения средств массовой коммуникации. М.: Ин-т социологии РАН, 2001. 202 с.
11. *Шалак В.И.* Мониторинг содержания и аудитории СМИ как средство диагностики состояния общества и необходимая предпосылка управления общественными процессами: Материалы междунар. научно-практ. конф. «Современные психотехнологии в образоовании, бизнесе, политике». Москва, 28 февр. – 2 марта 2001 г.
12. *Шалак В.И.* Контент-мониторинг текстовой информации: Материалы конф. «Проблемы обработки больших массивов неструктурированных текстовых документов». Москва, 23 мая 2001 г.
13. *Шалак В.И.* Математические методы компьютерного контент-анализа текстов // Тр. научно-исслед. семинара Логического центра Ин-та философии РАН. Вып. XVI. М., 2002.
14. *Шалак В.И.* Об использовании логики в контент-анализе: Тез. // Смирновские чтения. 4 Междунар. конф. М., 2003.

15. *Шалак В.И.* Современный контент-анализ: приложения в области политологии, социологии, психологии, культурологии, экономики и рекламы. М.: Омега-Л, 2004.
16. *Шалак В.И.* Проблемы логического анализа сети Интернет // Тез. докл. и выступлений IV Рос. филос. конгр. (Москва, 24-28 мая 2005 г.). М., 2005.
17. *Шалак В.И.* Публикации на страницах Интернет-сайта http:www.vaal.ru
18. *Bachus F.* Lp, a logic for representing and reasoning with statistical knowledge // Computational Intelligence. 1990. Vol. 6.
19. *Bachus F.* Probabilistic Belief Logics // http://www.cs.toronto.edu/~fbacchus/.
20. *Cox R.T.* Probability, Frequency and Reasonable Expectation // American J. of Physics/ 1946. Vol. 14, № 1.
21. *Fagin R., Halpern J.Y., Meggido N.* A logic for reasoning about probabilities // Information and Computation. Vol. 87(1-2).
22. *Gottwald S.* Axiomatizations of t-norm based logics – A survey // Soft Computing. 2000. Vol. 4.
23. *Gottwald S.* Many-valued Logic // Stanford Encyclopedia of Philosophy, 2004.
24. *Hajek A.* Interpretations of Probability // Stanford Encyclopedia of Philosophy, 2003.
25. *Hajek P.* Fuzzy Logic // Stanford Encyclopedia of Philosophy, 2002.
26. *Hajek P., Godo L., Esteva F.* A complete many-valued logic with product-conjunction // Archive for Mathematical Logic. Vol. 35, № 3.
27. *Hailperin T.* Probability Logic // Notre Dame J. of Logic. 1984. Vol. 25, № 3.
28. *Halpern J., Rabin M.* A Logic to Reason about Likelihood // Artificial Intelligence. 1987. Vol. 32.
29. *Halpern J.* An analysis of first-order logics of probability // Artificial Intelligence. 1990. Vol. 46.
30. *Halpern J., Koller D.* A Logic for Approximate Reasoning // Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning. 1992.
31. *Hart W.D.* Probability as Degree of Possibility // Notre Dame J. of Formal Logic. 1972. Vol. XIII, № 2.
32. *Hawthorne J.* Inductive Logic // Stanford Encyclopedia of Philosophy. 2005.
33. *Jaeger M.* A Logic for Inductive Probabilistic Reasoning.

34. *Janes E.T.* Probability Theory: The Logic of Science. 2002. http://bayes.wustl.edu/etj/etj.html.
35. *Kooi B.P.* Intensional and Statistical Probability. 2003.
36. *Miller D.* How Does Probability Theory Generalize Logic.
37. *Pfeifer N., Kleiter G.D.* Syllogistic reasoning with intermediate quantifiers. Technical report, 2005.
38. *Pfeifer N.* Contemporary syllogistics: comparative and quantitative syllogisms // *Kreuzbauer G. & Dorn G.* (Ed.). Argumentation in Theorie und Praxis: Philosophie und Didaktik des Argumentierens. Wien: LIT.
39. *Pfeifer N., Kleiter G.* Inference in Conditional Probability Logic // Proceedings of the 8th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty. Trest (Czech Republic), 2005.
40. *Pfeifer N., Kleiter G.* Towards a mental probability logic // Psychologica Belgica. 2005. Vol. 45(1).
41. *Sobel J.H.* Modus Ponens and Modus Tollens for Conditional Probabilities and Updating on Uncertain Evidence // http://www.scar.utoronto.ca/~sobel/.
42. *Terwijn S.A.* Probabilistic Logic and Induction // J. of Logic and Computation. 2005. Vol. 15(4).
43. *Talbott W.* Bayesian Epistemology // Stanford Encyclopedia of Philosophy, 2001.
44. *Wagner C.* Modus Tollens probabilized // British J. of Philosophy of Science. 2004. Vol. 55.
45. *Weatherson B.* Uncertainty Probability and Non-Classical Logic // Formal Epistemology Workshop, May 2004. Berkeley CA.
46. *Wheeler G.* Rational Acceptance and Conjunctive/Disjunctive Absorption // J. of Logic, Language and Information. 2005.

Научное издание

Шалак Владимир Иванович

Логический анализ сети Интернет

Утверждено к печати Ученым советом
Института философии РАН

В авторской редакции

Технический редактор *А.В. Сафонова*
Корректура автора

Лицензия ЛР № 020831 от 12.10.98 г.

Подписано в печать с оригинал-макета 20.12.05.
Формат 60х84 1/16. Печать офсетная. Гарнитура Таймс.
Усл. печ. л. 6,06. Уч.-изд. л. 3,31. Тираж 500 экз. Заказ № 052.

Оригинал-макет изготовлен в Институте философии РАН
Компьютерный набор и верстка автора

Отпечатано в ЦОП Института философии РАН
119992, Москва, Волхонка, 14